How to Estimate Change from Samples

Edith Cohen[*][†]   Haim Kaplan [†]

Abstract

Measurements, snapshots of a system, traffic matrices, and activity logs are typically collected repeatedly. *Difference queries* are then used to detect and localize changes for anomaly detection, monitoring, and planning. When the data is sampled, as is often done to meet resource constraints, queries are processed over the sampled data. We are not aware, however, of previously known estimators for L_p (p-norm) distances which are accurate when only a small fraction of the data is sampled.

We derive estimators for L_p distances that are nonnegative and variance optimal in a Pareto sense, building on our recent work on estimating general functions. Our estimators are applicable both when samples are independent or coordinated. For coordinated samples we present two estimators that tradeoff variance according to similarity of the data. Moreover, one of the estimators has the property that for all data, has variance is close to the minimum possible for that data.

We study performance of our Manhattan and Euclidean distance ($p = 1, 2$) estimators on diverse datasets, demonstrating scalability and accuracy – we obtain accurate estimates even when a small fraction of the data is sampled. We also demonstrate the benefit of tailoring the estimator to characteristics of the dataset.

1   Introduction

Data is commonly generated or collected repeatedly, where each *instance* has the form of a value assignment to a set of *keys*: Daily summaries of the number of queries containing certain keywords, transmitted bytes for IP flow keys, performance parameters (delay, throughput, or loss) for IP source destination pairs, environmental measurements for sensor locations, and requests for resources. In these examples, each set of values (instance) corresponds to a particular time or location. The universe of possible key values is fixed across instances but the values of a key are different in different instances.

Difference queries between instances facilitate anomaly detection, monitoring, and planning by detecting, measuring, and localizing change [6, 24]. Figure 1 shows two instances and difference measures that include the Euclidean and Manhattan norms.

keys:	1	2	3	4	5	6
Instance 1	5	0	4	5	8	7
Instance 2	7	10	3	0	6	7

RG_1	2	10	1	5	2	0
RG_{1+}	2	10	0	0	0	0
RG_{1-}	0	0	1	5	2	0
RG_2	4	100	1	25	4	0
RG_{2+}	4	100	0	0	0	0
RG_{2-}	0	0	1	25	4	0

Difference between instances for keys H:

$$L_{p\pm}^{p}(H) = \sum_{h \in H} \mathrm{RG}_{p\pm}(h)$$

$$L_{p\pm}(H) = \left(\sum_{h \in H} \mathrm{RG}_{p\pm}(h)\right)^{\frac{1}{p}}$$

diff: $\mathrm{RG}_p(h) = |v_1(h) - v_2(h)|^p$
up-only: $\mathrm{RG}_{p+}(h) = \max\{0, v_1(h) - v_2(h)\}^p$
down-only: $\mathrm{RG}_{p-}(h) = \max\{0, v_2(h) - v_1(h)\}^p$

Example Queries:
$L_1([6]) = 20$
$L_1(\{4, 5, 6\}) = 7$
$L_2^2(\{1, 5\}) = 32$
$L_2([6]) = \sqrt{134} \approx 11.6$
$L_{2-}([6]) = \sqrt{30} \approx 5.5$
$L_{2+}([6]) = \sqrt{104} \approx 10.2$

Figure 1: Difference measures between two instances on subset H of the keys. Top table shows values $v_i(h)$ for key $h \in [6]$ in instance $i \in [2]$. Bottom table shows single-key diffs.

Data collection and warehousing is subject to limitations on storage, throughput, and energy required for transmission. Even when the data is stored in full, exact processing of queries may be slow and resource consuming. Random

---

[*]Microsoft Research, SVC, USA

[†]The Blavatnik School of Computer Science, Tel Aviv University, Israel.

sampling of datasets is widely used as a means to obtain a flexible summary over which we can query the data while meeting these limitations [25, 33, 4, 3, 7, 19, 20, 1, 16, 1, 21, 10, 18, 8, 11].

Quality estimators are essential to scalable and accurate querying of sampled data. We seek estimators that are accurate when a small fraction of the data is sampled and are efficient to compute. Since differences are nonnegative, we are interested in nonnegative estimators.

A sampling algorithm maps data values and a set of random bits to a set of sampled keys. We focus on *weighted sampling*, meaning that the probability a key is sampled depends on its value. When these values are skewed, sampling schemes which favors heavier keys allow for more accurate estimates. When values are 0/1, keys with 0 value, which are often the majority of possible keys, have 0 inclusion probability and hence need not be explicitly considered – In an IP router, only a small subset of all possible keys IP addresses or IP flow keys are active.

As instances are dispersed in time or location, for scalability, the sample of one instance can not depend on values assumed in another [15, 12], but random bits can be public (when generated using random hash functions). The two extremes of the joint distribution of samples of different instances are *independent sampling* (independent sets of random bits for each instance) and *coordinated sampling* (identical sets of random bits). With coordinated sampling, similar instances have similar samples whereas independent samples of identical instances can be completely disjoint. These two sampling schemes have different strengths and therefore we consider both: While coordination [2, 31, 28, 30, 4, 3, 7, 19, 20, 1, 10, 21, 11, 15] allows for tighter estimates of many basic queries including distinct counts (set union), quantile sums, and as we shall see, difference norms, [7, 19, 20, 11, 15, 12], it also has pitfalls: (i) it results in unbalanced "burden" where same keys tend to be sampled across instances – an issue when, for example, sampling is performed prior to transmission to save power and (ii) variance on some queries – notably linear combinations of single-instance queries – is larger than with independent sampling ("total traffic on Monday-Wednesday", from daily summaries) – an issue if sample is primarily used for such queries.

Contribution:

We derive unbiased nonnegative estimators for L_p^p and the downward-only and upward-only components L_{p+}^p and L_{p-}^p. The sampling of each instance can be Poisson or bottom-k and samples of different instances can be independent or coordinated.

We estimate L_p^p as a sum over selected keys of nonnegative unbiased variance optimal RG_p estimates of the values assumed by the key (see Figure 1 for examples and definitions). Variance optimality is in a Pareto sense – another estimator with strictly lower variance on some data must have strictly higher variance on another data. L_p is estimated by the pth root of our L_p^p estimate. The bias of the L_p estimate decreases with the coefficient of variation of the (unbiased) L_p^p estimator, which decreases when more keys are selected. Similarly, the downward-only and upward-only components are estimated as respective sums of RG_{p-} and RG_{p+} estimates.

Over independently-sampled instances, we derive an optimal monotone estimator for RG_p of two values ($p > 0$). Monotonicity means that the estimate value is non-decreasing with the information we can glean from the outcome. Our construction adapts a technique we developed in [12].

Over coordinated samples, we apply [14, 13] to derive the L and U estimators which are unbiased, nonnegative, and variance-optimal. The L estimator has lower variance for data with small difference (range) whereas the U estimator performs better when the range is large. The L estimator is monotone and "variance competitive": on all data vectors, its variance is not too far off the minimum possible variance for the vector by a nonnegative unbiased estimator.

For $p = 1, 2$, we compute closed form expressions of estimators and their variance and also tight bounds on "competitiveness" of the L estimator. We evaluate and compare the performance of our L_1 and L_2^2 difference estimators on diverse data sets, which vary in size and magnitude of change, and relate observed performance to properties of the data. We also examine the behavior of the L and U estimators and provide guidelines to choosing between them based on properties of the data.

Roadmap: Section 3 contains necessary background and definitions. We present difference estimators for independent samples in Section 4 and for coordinated samples in Section 5. Section 6 contains an experimental evaluation.

2 Related work

There was little prior work on estimating difference norms from samples. This is at least partly because, under common schemes such as when sampling via random accesses, there are strong lower bounds [5, 12] on estimation quality, showing that most entries need to be sampled in order to obtain estimates with meaningful accuracy.

Our estimators, which are accurate even when only a small fraction of the data is sampled, critically depend on reproducibility of the "random bits" used by the sampling algorithm. More precisely, the inclusion probability of a key depends both on its value and a "random seed." Knowledge of the seed (which can be hash based) provides the estimator with additional power, since when a key is not sampled we are able to bound its value.

Fortunately, known seeds can be integrated with basic sampling schemes when data entries can be individually examined by the sampling algorithm, which is commonly the case when samples are produced as summaries of large data sets.

One can attempt to obtain nonnegative and unbiased estimates via classic inverse probabilities (Horvitz Thompson [23]): When the outcome reveals the value of the estimated quantity, the estimate is equal to the value divided by the probability of such an outcome. The estimate is 0 otherwise. Inverse probability estimates, however, are inapplicable to difference estimation over weighted samples, since they require that there is positive probability for outcomes that reveal the exact value of the estimated quantity. With multiple instances and weighted sampling, keys that have zero value in one instance and positive value in another have positive contribution to the difference but because zero values are never sampled, there is zero probability for determining the value from the outcome.

The only pre-existing satisfactory difference estimator we are aware of is for L_1 over coordinated samples, which uses the relation $|v_1 - v_2| = \max\{v_1, v_2\} - \min\{v_1, v_2\}$ to obtain an indirect estimate as the difference of two inverse probability estimates for the maximum and minimum [15]. Our U estimator for $p = 1$ is a strengthening of this L_1 estimator.

Lastly, difference estimation of streams was extensively studied using sketches of the streams (e.g. [17]), which are synopses that are not sample-based. With sketches it is possible to obtain tighter estimates on the difference between *complete* streams but sketches have limited usefulness for other queries and do not preserve information on values of particular keys, and in particular, do not naturally support subset queries.

3 Preliminaries

We denote by $v_i(h) \in V$ the value of key $h \in K$ in instance $i \in [r]$ and by the vector $\boldsymbol{v}(h)$, the values of key h in all instances. The *exponentiated range* function over values of a single key $\boldsymbol{v}(h)$ is:

$$\mathrm{RG}_p(\boldsymbol{v}) = (\max(\boldsymbol{v}) - \min(\boldsymbol{v}))^p \quad (p > 0) \tag{1}$$

where $\max(\boldsymbol{v}) \equiv \max_i v_i$ and $\min(\boldsymbol{v}) = \min_i v_i$ are the maximum and minimum entry values of the vector $\boldsymbol{v}$. We omit the subscript when $p = 1$. For two instances ($r = 2$) we use the following to isolate upward-only and downward-only changes:

$$\mathrm{RG}_{p+}(v_1, v_2) = \max\{v_1 - v_2, 0\}^p$$
$$\mathrm{RG}_{p-}(v_1, v_2) = \max\{v_2 - v_1, 0\}^p$$

For a selected subset $H \subset K$ of keys, we define

$$L_p^p(H) = \sum_{h \in H} \mathrm{RG}_p(\boldsymbol{v}(h)) . \tag{2}$$

The p-norm of the difference of two instances ($r = 2$) is $L_p(H) \equiv (L_p^p(H))^{1/p} = ||v_1(H) - v_2(H)||_p$. For upward-only and downward-only change we use $L_{p\pm}^p(H) = \sum_{h \in H} \mathrm{RG}_{p\pm}(v_1(h), v_2(h))$.

When data is sampled, we estimate L_p^p, L_{p+}^p, and L_{p-}^p, by summing estimates for the respective single-key primitives (RG_p, RG_{p+} and RG_{p-}) over keys in H. We use unbiased estimators for the primitives, which result, from

	1	2	3	4	5	6
Instance 1	5	0	4	5	8	7
Instance 2	7	10	3	0	6	7

Independent sampling, PPS						
seeds u_1	0.23	0.29	0.84	0.15	0.58	0.19
seeds u_2	0.81	0.17	0.48	0.36	0.15	0.49
v_1/u_1	21.7	0	4.8	33.3	13.8	36.8
v_2/u_2	8.6	58.8	6.25	0	41.7	14.3

Coordinated sampling, PPS						
seeds u	0.23	0.29	0.84	0.15	0.58	0.19
v_1/u	21.7	0	4.8	33.3	13.8	36.8
v_2/u	30.4	34.5	3.6	0	10.3	36.8

Poisson samples $E[|S|] = 3$:

	τ_i	S
independent		
1	29/3	$\{1, 4, 5, 6\}$
2	11	$\{2, 5, 6\}$
coordinated $u \equiv u_1$		
1	29/3	$\{1, 4, 5, 6\}$
2	11	$\{1, 2, 5, 6\}$

priority samples $|S| = 3$:

	eff. τ_i	S
independent		
1	13.8	$\{4, 5, 6\}$
2	8.6	$\{2, 5, 6\}$
coordinated $u \equiv u_1$		
1	13.8	$\{4, 5, 6\}$
2	10.3	$\{1, 2, 6\}$

Figure 2: Independent and coordinated samples of two instances. Seed selection and ratios. Poisson PPS samples of expected size 3 and priority samples of size 3.

linearity of expectation, in unbiased estimates for the sums. Since only a fraction of keys is sampled, our estimates for each primitive generally have high CV (coefficient of variation, which is the ratio of the square root of the variance to the mean). Since estimates for different keys are (pairwise) independent (or nonpositively correlated), variance is (sub)-additive and the CV decreases when $|H|$ increases, allowing for accurate estimates when H is sufficiently large. Unbiasedness of the single-key estimators is essential here. Since unbiasedness is not preserved under exponentiations, we need to carefully tailor to the exponent value p.

We estimate $L_p(H)$ by taking the pth root of the estimate for $L_p^p(H)$. This estimate is biased, but the error is small when the CV of our $L_p^p(H)$ estimate is small.

Sampling scheme of instances. Our estimators apply to Poisson sampling, where keys are sampled independently, and bottom-k (order) sampling, that yields a sample size of exactly k. Bottom-k sampling includes Priority (Sequential Poisson) sampling and weighted sampling without replacement [30, 29, 27, 9, 18, 10, 11]. We state these classic schemes in a way which allows the "random" bits to be reproducible, first by the sampling algorithm, to facilitate coordination, and also by the estimator, to yield strong estimates.

We reuse notation from [12, 14, 13]. Sampling is specified by a set of nondecreasing continuous functions τ_i^h, defined on the interval $[0, 1]$. Each key h in instance i is associated with a random *seed* value $u_i(h) \sim U[0, 1]$ chosen uniformly at random. To make randomization reproducible, $u_i(h)$ is generated via a random hash function (pairwise independence and fewer bits in the representation of the seed suffice, but we skip these details here). With Poisson sampling,

$$h \text{ is sampled in instance } i \iff v_i(h) \geq \tau_i^h(u_i(h)) . \tag{3}$$

A bottom-k sample of instance i includes the k keys with largest ratios $r_i(h) = v_i(h)/\tau_i^h(u_i(h))$. Samples of different instances are *independent* when the seeds $u_i(h)$ are independent for all i. They are *coordinated* (shared-seed) if the same seed is used for the same key in all instances, that is, $\forall h \in K, \forall i \in [r], u_i(h) = u_1(h) \equiv u(h)$.

When threshold functions have the form $\tau_i^h(x) = ax$ (to simplify notation we use $\tau_i^h(x) \equiv \tau_i^h x$, treating τ_i^h as constant), the corresponding Poisson samples are *PPS* (Probability Proportional to Size) [22]. Strictly, PPS sampling assumes that τ_i^h are consistent across the instance, that is, $\tau_i^h \equiv \tau_i$, but our analysis is general.

Sampling can be performed efficiently both when the threshold τ_i is fixed or when set adaptively by a streaming algorithm to achieve a specified expected sample size $E[|S|] = \sum_{h \in K} \min\{1, v_i(h)/\tau_i\}$. As an example, to obtain a PPS Poisson sample of expected size $E[|S|] = 3$ for the instances in Figure 1 we use $\tau_1 = 29/3$ (instance 1) and $\tau_2 = 33/3 = 11$ (instance 2).

The bottom-k sample obtained with $\tau_i^h(x) \equiv x$ is a priority (sequential Poisson) sample [27, 18, 32]. Weighted sampling without-replacement is obtained with thresholds $\tau_i^h(x) \equiv -\ln x$. Figure 2 shows PPS and priority samples obtained with respect to random seeds for the two instances in Figure 1.

Sampling model (single key): The exponentiated range estimators are applied to samples of the same key h across instances $i \in r$. That is, we work with the restrivtion of the sample to one key at a time.

With Poisson sampling, for key h, we can obtain from the sample (3), the values of sampled entries of key h. The seed vector $\boldsymbol{u} \equiv \boldsymbol{u}(h)$ and the thresholds $\boldsymbol{\tau} = (\tau_1(h), \ldots, \tau_r(h))$ are all available to the estimator. With Bottom-k

sampling of instances, the threshold is not readily available, so we work with *effective* thresholds as follows. We condition the inclusion of h on seeds of other keys being fixed [10, 18] and define $\tau_i^h \equiv \tau_i$ to be the inverse of the kth largest $r_i(h)$ of keys in instance i with h excluded (which is the $k + 1$st largest ratio over all keys in the instance). From here onward, we omit from the notation the reference to the key h and focus on exponentiated range estimators. We return to sum aggregates only for the experiments in Section 6.

The data (values of a single key h in instances $i \in [r]$) is $\boldsymbol{v} = (v_1, v_2, \ldots, v_r) \in \mathbf{V} = V^r$ (we mostly assume $\mathbf{V} \subset \mathbb{R}_{\geq 0}^r$). The outcome S depends on the data $\boldsymbol{v}$, random seed vector $\boldsymbol{u}$ and threshold functions $\boldsymbol{\tau}$. The ith entry is included in S if and only if its value is at least $\tau_i(u_i)$:

$$i \in S \iff v_i \geq \tau_i(u_i) \,.$$

The set of all data vectors consistent with outcome S (we treat $\boldsymbol{u} \in [0,1]^r$ as included with S) is

$$\begin{aligned} V^*(S) &= \{\boldsymbol{v} \in \mathbf{V} \mid S = S(\boldsymbol{u}, \boldsymbol{v})\} \\ &= \{\boldsymbol{z} \mid \forall i \in [r], i \in S \wedge z_i = v_i \ \vee \ i \notin S \wedge z_i < \tau_i(u_i)\} \,. \end{aligned}$$

We can equivalently define the outcome as the set $V^*(S)$ since it captures all information available to the estimator.

Estimators: An *estimator* $\hat{f}$ for $f : \mathbf{V}$ is a numeric function applied to the outcome. To be well defined in continuous domains, $\hat{f}$ should be (Lebesgue) integrable. For exponentiated ranges, which are nonnegative quantities, we are interested in estimators that are *nonnegative* $\hat{f}(S) \geq 0$ for all S. As explained earlier, since we sum many estimates, we would like each estimate to be *unbiased* $\mathsf{E}[\hat{f}|\boldsymbol{v}] = f(\boldsymbol{v})$. Other properties we seek are *bounded variance* on all data, and *variance-optimality* (respectively, variance$^+$-optimality): there is no (resp., nonnegative) estimator with same or lower variance on all data and strictly lower on some data. An intuitive property that is sometimes desirable is *monotonicity*: the estimate value is non decreasing with the information on the data that we can glean from the outcome $V^*(S) \subset V^*(S') \implies \hat{f}(S) \geq \hat{f}(S')$.

Order-based variance optimality: Given a partial order $\prec$ on $\mathbf{V}$, an estimator $\hat{f}$ is $\prec$-optimal (respectively, $\prec^+$-optimal) if it is unbiased (resp., nonnegative) and for all data $\boldsymbol{v}$, minimizes variance for $\boldsymbol{v}$ conditioned on the variance being minimized for all preceding vectors. Formally, if there is no other unbiased (resp., nonnegative) estimator that has strictly lower variance on some data $\boldsymbol{v}$ and at most the variance of $\hat{f}$ on all vectors that precede $\boldsymbol{v}$. Order-based optimality implies variance optimality.

4 Independent PPS sampling

The outcome $S(\boldsymbol{u}, \boldsymbol{v})$ is determined by the data $\boldsymbol{v}$ and a random seed vector $\boldsymbol{u} \in [0,1]^r$ with independent entry values.

$$V^*(S) = \{\boldsymbol{z} \mid \forall i \in [r], i \in S \wedge z_i = v_i \ \vee \ i \notin S \wedge z_i < \tau_i u_i\} \,.$$

We derive the L estimator, $\hat{\mathrm{RG}}_p^{(L)}$, which is the unique symmetric, monotone, and variance$^+$ optimal estimator, by applying our framework from [12] to construct the estimator for a function f. The application has two ingredients: The first is a method to construct a $\prec$-optimal estimator $\hat{f}^{(\prec)}$ for with respect to a partial order $\prec$ on data vectors. The second ingredient is to identify a partial order $\prec$ so that the estimator $\hat{f}^{(\prec)}$ is nonnegative, and therefore, $\prec^+$-optimal.

Review of the construction of $\hat{f}^{(\prec)}$. The *determining vector* $\phi(S)$ of an outcome S is a $\prec$-minimal vector in the closure of consistent vectors: $\phi(S) = \min_{\prec} \mathrm{cl}(V^*(S))$. Accordingly, we can specify the sets $\phi^{-1}(\boldsymbol{v})$ of all outcomes determined by $\boldsymbol{v}$ and all outcomes $\mathcal{S}_0(\boldsymbol{v})$ that precede $\boldsymbol{v}$, that is, consistent with $\boldsymbol{v}$ but determined by a vector that precedes $\boldsymbol{v}$:

$$\begin{aligned} \phi^{-1}(\boldsymbol{v}) &= \{S | \boldsymbol{v} = \phi(S)\} \\ \mathcal{S}_0(\boldsymbol{v}) &= \{S | \boldsymbol{v} \in V^*(S) \wedge \phi(S) \prec \boldsymbol{v}\} \end{aligned}$$

The estimator $\hat{f}^{(\prec)}$ is the same for all outcomes with same determining vector, and therefore we can specify it in terms of the determining vector $\hat{f}^{(\prec)}(S) \equiv \hat{f}^{(\prec)}(\phi(S))$. We now state constraints that must be satisfied by $\hat{f}^{(\prec)}$. The

5

contribution of the outcomes $\mathcal{S}_0(\boldsymbol{v})$ to the expectation $\mathsf{E}[\hat{f}^{(\prec)}|\boldsymbol{v}]$ is

$$f_0(\boldsymbol{v}) = \mathsf{E}[\hat{f}^{(\prec)}|\mathcal{S}_0(\boldsymbol{v}), \boldsymbol{v}]\mathrm{PR}[\mathcal{S}_0(\boldsymbol{v})|\boldsymbol{v}] \;,$$

where $\mathsf{E}[\hat{f}^{(\prec)}|\mathcal{S}_0(\boldsymbol{v}), \boldsymbol{v}]$ is the expectation of $\hat{f}^{(\prec)}$ on outcomes that precede $\boldsymbol{v}$ and $\mathrm{PR}[\mathcal{S}_0(\boldsymbol{v})|\boldsymbol{v}]$ is the probability that the outcome precedes $\boldsymbol{v}$ when the data is $\boldsymbol{v}$. For all vectors $\boldsymbol{v} \in V$, we require (this is necessary for unbiasedness) that

$$\textbf{If } \mathrm{PR}[\phi^{-1}(\boldsymbol{v})|\boldsymbol{v}] = 0 \textbf{ then } f_0(\boldsymbol{v}) \equiv f(\boldsymbol{v}). \tag{4}$$

$$\textbf{Else, } \hat{f}^{(\prec)}(\boldsymbol{v}) = \frac{f(\boldsymbol{v}) - f_0(\mathbf{v})}{\mathrm{PR}[\phi^{-1}(\boldsymbol{v})|\boldsymbol{v}]} \tag{5}$$

where $\mathrm{PR}[\phi^{-1}(\boldsymbol{v})|\boldsymbol{v}]$ is the probability that the outcome is determined by $\boldsymbol{v}$ when the data vector is $\boldsymbol{v}$.

Choice of $\prec$. For RG_p, we choose $\prec$ so that the relation between vectors is according to an increasing lexicographic order on the lists $L(\boldsymbol{v})$, which we define to be the sorted multiset of differences $\{v_i - \min(\boldsymbol{v}) \mid i \in [r]\}$. The $\prec$-minimum vectors are those with all entries having equal values. With our choice of $\prec$, the determining vector $\phi(S)$ is as follows: if $S = \emptyset$ (no entries are sampled), $\phi(S) = \boldsymbol{0}$. Otherwise, if $h \in S$ then $\phi(S)_h = v_h$ and if $h \notin S$ then $\phi(S)_h = \min\{\min_{i \in S} v_i, \min_{i \notin S} u_i\tau_i\}$. The mapping of outcomes to determining vectors for $r = 2$ is shown in Table 1 (Right).

$\boldsymbol{v} = (v_1, v_2)$	$\hat{\mathrm{RG}}_p^{(L)}(\boldsymbol{v})$
$\boldsymbol{v} = (0,0)$	0
$v_1 \geq v_2 > \tau_2$	$\frac{\tau_1}{\min\{\tau_1, v_1\}}(v_1 - v_2)^p$
$v_1 \geq v_2 \leq \tau_2$	$\frac{p\tau_1\tau_2}{\min\{v_1, \tau_1\}} \int_{\max\{0, v_1 - \tau_2\}}^{v_1 - v_2} \frac{y^{p-1}}{v_1 - y}dy +$ $+ \frac{\tau_1 \max\{0, v_1 - \tau_2\}^p}{\min\{v_1, \tau_1\}}$

outcome S		$\phi(S)_1$	$\phi(S)_2$
$S = \emptyset$	:	0	0
$S = \{1\}$	:	v_1	$\min\{u_2\tau_2, v_1\}$
$S = \{2\}$	:	$\min\{u_1\tau_1, v_2\}$	v_2
$S = \{1, 2\}$	:	v_1	v_2

Table 1: Left: Estimator $\hat{\mathrm{RG}}_p^{(L)}$ for $p > 0$ and $r = 2$ over independent samples, stated as a function of the determining vector (v_1, v_2) when $v_1 \geq v_2$ (case $v_2 > v_1$ is symmetric). Right: mapping of outcomes to determining vectors.

Derivation of $\hat{\mathrm{RG}}_p^{(L)}$. To obtain the estimator, we solve (5) for all $\boldsymbol{v}$ such that $\mathrm{PR}[\phi^{-1}(\boldsymbol{v})|\boldsymbol{v}] > 0$ and verify that $f_0(\boldsymbol{v}) \equiv f(\boldsymbol{v})$ for all $\boldsymbol{v}$ such that $\mathrm{PR}[\phi^{-1}(\boldsymbol{v})|\boldsymbol{v}] = 0$. The estimator $\hat{\mathrm{RG}}_p^{(L)}$ ($p > 0$) when $r = 2$ is specified in Table 1 through a mapping of determining vectors to estimate values. We can verify that for all $p > 0$, the estimator $\hat{\mathrm{RG}}_p^{(L)}$ is nonnegative, monotone ($\hat{\mathrm{RG}}_p^{(L)}(v, x)$ is non-increasing for $x \in (0, v]$) and has finite variances (follow from $\int_0^v \hat{\mathrm{RG}}_p^{(L)}(v, x)^2 dx < \infty$). Table 2 shows explicit expressions of $\hat{\mathrm{RG}}^{(L)}$ and $\hat{\mathrm{RG}}_2^{(L)}$.

Vectors with $\mathrm{PR}[\phi^{-1}(S)] = 0$ are exactly those with one positive and one zero entry and we can verify that (4) is satisfied, that is, that $\hat{\mathrm{RG}}_p^{(L)}$ ($p > 0$) is unbiased on these vectors. We conjecture that a solution of (5) for $r > 2$ is also nonnegative, and monotone.

$\boldsymbol{v} = (v_1, v_2)$	$\hat{\mathrm{RG}}^{(L)}(\boldsymbol{v})$
$\boldsymbol{v} = (0,0)$	0
$v_1 \geq v_2 > \tau_2$	$\frac{\tau_1}{\min\{\tau_1, v_1\}}(v_1 - v_2)$
$v_1 \geq v_2 \leq \tau_2$	$\frac{\tau_1\tau_2}{\min\{\tau_1, v_1\}} \ln\left(\frac{\min\{v_1, \tau_2\}}{v_2}\right) + \frac{\tau_1 \max\{0, v_1 - \tau_2\}}{\min\{v_1, \tau_1\}}$

$\boldsymbol{v} = (v_1, v_2)$	$\hat{\mathrm{RG}}_2^{(L)}(\boldsymbol{v})$
$\boldsymbol{v} = (0,0)$	0
$v_1 \geq v_2 > \tau_2$	$\frac{\tau_1}{\min\{\tau_1, v_1\}}(v_1 - v_2)^2$
$v_1 \geq v_2 \leq \tau_2$	$\frac{2\tau_1\tau_2}{\min\{\tau_1, v_1\}}\left(v_2 - \min\{v_1, \tau_2\} + v_1 \ln \frac{\min\{v_1, \tau_2\}}{v_2}\right)$ $+ \frac{\tau_1 \max\{0, v_1 - \tau_2\}^2}{\min\{v_1, \tau_1\}}$

Table 2: Explicit form of estimators $\hat{\mathrm{RG}}^{(L)}$ and $\hat{\mathrm{RG}}_2^{(L)}$ for $r = 2$ over independent samples. Estimator is stated as a function of the determining vector (v_1, v_2) when $v_1 \geq v_2$ (case $v_2 \geq v_1$ is symmetric).

We now provide details on the derivation. We consider vectors $\boldsymbol{v}$ in increasing $\prec$ order and solve (5) for $f^{(\prec)}$ on outcomes with determining vector $\boldsymbol{v}$.

We express $\hat{\mathrm{RG}}_p^{(L)}(v, v - \Delta)$ ($p > 0$, $\Delta \in [0, v]$) as a function of τ_1 and τ_2.

• **Case:** $v - \Delta \geq \tau_2$. Estimate can be positive only when $u_1\tau_1 \leq v$, which happens with probability $\min\{1, v/\tau_1\}$. We solve the equality $\Delta^p = \min\{1, v/\tau_1\}\hat{\text{RG}}_p^{(L)}$, obtaining

$$\hat{\text{RG}}_p^{(L)}(v, v - \Delta) = \frac{\tau_1}{\min\{v, \tau_1\}}\Delta^p . \tag{6}$$

• **Case:** $v - \Delta < \tau_2$. From (5):

$$\Delta^p = \frac{\min\{v, \tau_1\}}{\tau_1}\frac{v - \Delta}{\tau_2}\hat{\text{RG}}_p^{(L)}(v, v - \Delta) + \frac{\min\{v, \tau_1\}}{\tau_1\tau_2}\int_{\max\{0, v-\tau_2\}}^{\Delta}\hat{\text{RG}}_p^{(L)}(v, v - y)dy$$

Taking a partial derivative with respect to Δ, we obtain

$$\frac{\partial\hat{\text{RG}}_p^{(L)}(v, v - \Delta)}{\partial\Delta} = \frac{p\tau_1\tau_2}{\min\{v, \tau_1\}}\frac{\Delta^{p-1}}{v - \Delta}$$

We use the boundary value for $\Delta = \max\{0, v - \tau_2\}$:

$$\hat{\text{RG}}_p^{(L)}(v, \min\{v, \tau_2\}) = \frac{\tau_1}{\min\{v, \tau_1\}}\max\{0, v - \tau_2\}^p ,$$

and obtain

$$\hat{\text{RG}}_p^{(L)}(v, v - \Delta) = \frac{p\tau_1\tau_2}{\min\{v, \tau_1\}}\int_{\max\{0, v-\tau_2\}}^{\Delta}\frac{y^{p-1}}{v - y}dy + \frac{\tau_1\max\{0, v - \tau_2\}^p}{\min\{v, \tau_1\}}$$

The special case $\tau_1 = \tau_2 = \tau$: The estimators $\hat{\text{RG}}^{(L)}$ and $\hat{\text{RG}}_2^{(L)}$ as a function of the determining vector and their variance are provided in Tables 3 and 4. For data vectors where $v_1 \geq v_2 \geq \tau$, $\hat{\text{RG}}^{(L)} = v_1 - v_2$ and $\text{VAR}[\hat{\text{RG}}^{(L)}] = 0$. If $v_1 \geq \tau \geq v_2$, $\hat{\text{RG}}^{(L)} = \tau\ln\frac{\tau}{v_2} + v_1 - \tau$, and $\text{VAR}[\hat{\text{RG}}^{(L)}] = -2\tau v_2\ln(\frac{\tau}{v_2}) - v_2^2 + (\tau)^2$. Finally, if $v_2 \leq v_1 \leq \tau$, $\hat{\text{RG}}^{(L)}(v_1, v_2) = \frac{(\tau)^2}{v_1}\ln\frac{v_1}{v_2}$ and

$$\text{VAR}[\hat{\text{RG}}^{(L)}|(v_1, v_2)] =$$
$$= \frac{v_1 v_2}{(\tau)^2}(\frac{(\tau)^2}{v_1}\ln\frac{v_1}{v_2} - (v_1 - v_2))^2 + +(1 - \frac{v_1^2}{(\tau)^2})(v_1 - v_2)^2 + \frac{v_1}{(\tau)^2}\int_{v_2}^{v_1}(\frac{(\tau)^2}{v_1}\ln\frac{v_1}{y} - (v_1 - v_2))^2 dy$$
$$= 2(\tau)^2(1 - \frac{v_2}{v_1}\ln\frac{v_1}{v_2} - \frac{v_2}{v_1}) - (v_1 - v_2)^2 .$$

Determining vector $v_1 \geq v_2$	$\hat{\text{RG}}^{(L)}(v_1, v_2)$
$v_1 \geq v_2 \geq \tau$	$v_1 - v_2$
$v_1 \geq \tau \geq v_2$	$\tau\ln\frac{\tau}{v_2} + v_1 - \tau$
$v_2 \leq v_1 \leq \tau$	$\frac{(\tau)^2}{v_1}\ln\frac{v_1}{v_2}$

Data $v_1 \geq v_2$	$\text{VAR}[\hat{\text{RG}}^{(L)}|v]$
$v_1 \geq v_2 \geq \tau$	0
$v_1 \geq \tau \geq v_2$	$-2\tau v_2\ln(\frac{\tau}{v_2}) - v_2^2 + \tau^2$
$v_2 \leq v_1 \leq \tau$	$2\tau^2(1 - \frac{v_2}{v_1}\ln\frac{v_1}{v_2} - \frac{v_2}{v_1}) - (v_1 - v_2)^2$

Table 3: Estimator $\hat{\text{RG}}^{(L)}$ and its variance for independent samples.

Similarly, for $v_2 \leq v_1 \leq \tau$, $\hat{\text{RG}}_2^{(L)}(v_1, v_2) = 2(\tau)^2(\ln\frac{v_1}{v_2} - \frac{v_1 - v_2}{v_1})$. The variance when $v_1 \geq \tau$ is the same as for shared-seed sampling (see next section).

5 Shared-seed sampling

The outcome $S(u, v)$ is determined by the data v and a *scalar* seed value $u \in (0, 1]$, drawn uniformly at random: Entry i is included in S if and only if $v_i \geq \tau_i(u)$, where τ_i is a non-decreasing continuous function with range containing $(\min V, \max V)$.

Data $v_1 \geq v_2$	$\mathrm{VAR}[\hat{\mathrm{RG}}_2^{(L)}|\boldsymbol{v}]$
$v_1 \geq v_2 \geq \tau$	0
$v_1 \geq \tau \geq v_2$	$-4v_1v_2\tau(2v_1 - v_2)\ln\frac{\tau}{v_2}$
	$+4v_1v_2(\tau)^2 + \frac{(\tau)^4}{3} + \frac{8v_2^3\tau}{3}$
	$-6v_1v_2^2\tau - 4v_1^2 v_2^{\frac{3}{2}}$
	$-v_2^4 + 4v_1v_2^3 + 4v_1^2(\tau)^2 - 2v_1(\tau)^3$
$v_2 \leq v_1 \leq \tau$	$\frac{2(\tau)^2}{3v_1}\left(4v_2^3 + 5v_1^3 - 9v_1 v_2^2\right) - (v_1 - v_2)^4$
	$-4(\tau)^2(2v_1 - v_2)v_2\ln(\frac{v_1}{v_2})$

condition	$\hat{\mathrm{RG}}_2^{(L)}(v_1, v_2)$
$v_1 \geq v_2 \geq \tau$	$(v_1 - v_2)^2$
$v_1 \geq \tau \geq v_2$	$v_1^2 - (\tau)^2 - 2\tau(v_1 - v_2) + 2\tau v_1\ln\frac{\tau}{v_2}$
$v_2 < v_1 \leq \tau$	$2(\tau)^2(\ln\frac{v_1}{v_2} - \frac{v_1-v_2}{v_1})$

Table 4: Estimator $\hat{\mathrm{RG}}_2^{(L)}$ and its variance for independent samples.

Our RG_p ($p > 0$) estimators are derived by applying our general theory on shared-seed estimators [14, 13]. We derive two unbiased nonnegative variance$^+$-optimal estimators, the L estimator $\hat{\mathrm{RG}}_p^{(L)}$ and the U estimator $\hat{\mathrm{RG}}_p^{(U)}$. As a reference point, we also consider for each vector $\boldsymbol{v}$, the *$\boldsymbol{v}$-optimal estimate values* $\hat{\mathrm{RG}}_p^{(\boldsymbol{v})}$. An estimator is $\boldsymbol{v}$-optimal if amongst all estimators that are nonnegative and unbiased for all data, it has the minimum possible variance when the data is $\boldsymbol{v}$. It turns out that the values assumed by a $\boldsymbol{v}$-optimal estimator on outcomes consistent with $\boldsymbol{v}$ are unique, up to equivalence, and we refer to them as the $\boldsymbol{v}$-optimal estimates.

We compute closed form expressions of estimators and variances when $\boldsymbol{\tau}$ has all entries equal (to the scalar τ). The expressions for the upward-only and downward-only variants follow those for $\hat{\mathrm{RG}}_p$ and are omitted.

Structure of the set of outcomes. The set of data vectors consistent with outcome $S(u, \boldsymbol{v})$ is

$$V^*(S) = \{\boldsymbol{z} | \forall i \in [r], i \in S \wedge z_i = v_i \vee i \notin S \wedge z_i < \tau_i(u)\}.$$

From the outcome $S(u, \boldsymbol{v})$, we can determine not only $V^*(S(u, \boldsymbol{v}))$ but also $V^*(S(x, \boldsymbol{v}))$ for all $x \geq u$. Observe that the sets $V^*(S(u, \boldsymbol{z}))$ are the same for all consistent data vectors $\boldsymbol{z} \in V^*(S(u, \boldsymbol{v}))$. Fixing the data $\boldsymbol{v}$, the upper bounds $\tau_i(u)$ on values of entries that are not sampled are non-decreasing functions of u and therefore, the set $V^*(S(u, \boldsymbol{v}))$ is non-decreasing with u and the set of sampled entries is non-increasing. This means that the information on the data that we can glean from the outcome increases when u decreases.

The lower bound function. To proceed, we need to define the *lower bound function* $\underline{\mathrm{RG}}_p$:

$$\underline{\mathrm{RG}}_p(S) = \inf_{\boldsymbol{v} \in V^*(S)} \mathrm{RG}_p(\boldsymbol{v}),$$

which maps an outcome S to the infimum of RG_p values on vectors that are consistent with the outcome. For RG, the lower bound is the difference between a lower bound on the maximum entry and an upper bound on the minimum entry.

$$\underline{\mathrm{RG}}(S) = \max_{i \in S} v_i - \min\{\min_{i \in S} v_i, \min_{i \notin S} \tau_i(\rho)\}.$$

The lower bound on RG_p is the pth power of the respective bound on RG, that is, $\underline{\mathrm{RG}}_p(S) = \underline{\mathrm{RG}}(S)^p$. For $S(u, \boldsymbol{v})$, we use the notation $\underline{\mathrm{RG}}_p(S) \equiv \underline{\mathrm{RG}}_p^{(\boldsymbol{v})}(u)$ which is convenient when we fix $\boldsymbol{v}$ while varying the seed u.

For PPS sampling with all-entries-equal $\boldsymbol{\tau}$:

condition	$|S|$	$\underline{\mathrm{RG}}(S)$
$u > \frac{\max(\boldsymbol{v})}{\tau}$	0	0
$\frac{\max(\boldsymbol{v})}{\tau} \geq u \geq \frac{\min(\boldsymbol{v})}{\tau}$	$1 \dots r-1$	$\max(\boldsymbol{v}) - u\tau$
$u < \frac{\min(\boldsymbol{v})}{\tau}$	r	$\underline{\mathrm{RG}}(\boldsymbol{v})$

$\boldsymbol{v}$-optimality. For a data vector $\boldsymbol{v}$, we can determine when a nonnegative and unbiased estimator $\hat{\mathrm{RG}}_p$ has minimum variance on $\boldsymbol{v}$. We use the notation $H_{\mathrm{RG}_p}^{(\boldsymbol{v})}(u)$ for the lower boundary of the convex hull (lower hull) of $\underline{\mathrm{RG}}_p^{(\boldsymbol{v})}(u)$. This function is monotone non-increasing and therefore differentiable almost everywhere.

Theorem 5.1. *[14] A nonnegative unbiased estimator $\hat{\text{RG}}_p$ minimizes* $\text{VAR}[\hat{\text{RG}}_p|\boldsymbol{v}] \iff H^{(\boldsymbol{v})}_{\text{RG}_p}(u) = \int_u^1 \hat{\text{RG}}_p(\boldsymbol{v}, x) dx$ $\iff$ almost everywhere

$$\hat{\text{RG}}_p^{(\boldsymbol{v})}(u) = -\frac{dH^{(\boldsymbol{v})}_{\text{RG}_p}(u)}{du} . \tag{7}$$

Note that the specification (7) of the $\boldsymbol{v}$-optimal estimates on outcomes consistent with $\boldsymbol{v}$ is unique (in an almost everywhere sense). The estimates (7) are monotone non-increasing in u. Observe that the specification for different vectors with overlapping sets of consistent outcomes can be inconsistent and thus, it is not possible to obtain a single universally optimal estimator that is $\boldsymbol{v}$-optimal for all $\boldsymbol{v}$.

We can now specify $\hat{\text{RG}}_p^{(\boldsymbol{v})}$ for PPS sampling with all-entries-equal τ. The function $\underline{\text{RG}}_p^{(\boldsymbol{v})}(u)$ is $\max\{0, \max(\boldsymbol{v}) - \tau\}$ for $u \geq \frac{\max(\boldsymbol{v})}{\tau}$ and equal to $\text{RG}_p(\boldsymbol{v})$ for $u \leq \frac{\min(\boldsymbol{v})}{\tau}$. Therefore for $u \geq \frac{\max(\boldsymbol{v})}{\tau}$, the lower hull is $H^{(\boldsymbol{v})}_{\text{RG}_p}(u) = 0$ and $\hat{\text{RG}}_p^{(\boldsymbol{v})}(u) = 0$.

For $p \leq 1$, the function is concave for $u \in [\frac{\min(\boldsymbol{v})}{\tau}, \frac{\max(\boldsymbol{v})}{\tau}]$. The lower hull is therefore a linear function for $u \leq \frac{\max(\boldsymbol{v})}{\tau}$: when $\max(\boldsymbol{v}) \leq \tau$, $H^{(\boldsymbol{v})}_{\text{RG}_p}(u) = \text{RG}_p(\boldsymbol{v})(1 - u\frac{\tau}{\max(\boldsymbol{v})})$ and when $\max(\boldsymbol{v}) \geq \tau$, $H^{(\boldsymbol{v})}_{\text{RG}_p}(u) = \text{RG}_p(\boldsymbol{v}) - u(\text{RG}_p(\boldsymbol{v}) - (\max(\boldsymbol{v}) - \tau)^p)$. The $\boldsymbol{v}$-optimal estimator is therefore constant for $u \leq \min\{1, \frac{\max(\boldsymbol{v})}{\tau}\}$: $\hat{\text{RG}}_p^{(\boldsymbol{v})}(u) = \text{RG}(\boldsymbol{v})\frac{\tau}{\max(\boldsymbol{v})}$ when $\max(\boldsymbol{v}) \leq \tau$, and $\hat{\text{RG}}_p^{(\boldsymbol{v})}(u) = \text{RG}_p(\boldsymbol{v}) - (\max(\boldsymbol{v}) - \tau)^p$ when $\max(\boldsymbol{v}) \geq \tau$.

For $p > 1$, $\underline{\text{RG}}_p^{(\boldsymbol{v})}(u)$ is convex for $u \in [\frac{\min(\boldsymbol{v})}{\tau}, \frac{\max(\boldsymbol{v})}{\tau}]$. Geometrically, the lower hull follows the lower bound function for $u > \alpha$, where α is the point where the slope of the lower bound function is equal to the slope of a line segment connecting the current point to the point $(0, \text{RG}_p(\boldsymbol{v}))$. For $u \leq \alpha$, the lower hull follows this line segment and is linear. Formally, the point α is the solution of

$$\text{RG}_p(\boldsymbol{v}) = (\max(\boldsymbol{v}) - x\tau)^{p-1}(p\tau + \max(\boldsymbol{v}) - x\tau) .$$

If there is no solution $\alpha \in [\frac{\min(\boldsymbol{v})}{\tau}, \min\{1, \frac{\max(\boldsymbol{v})}{\tau}\}]$, we use $\alpha = \min\{1, \frac{\max(\boldsymbol{v})}{\tau}\}$. The estimates for $u \in [\alpha, \min\{1, \frac{\max(\boldsymbol{v})}{\tau}\}]$ are $\hat{\text{RG}}_p^{(\boldsymbol{v})}(u) = -\frac{d\underline{\text{RG}}_p^{(\boldsymbol{v})}(u)}{du} = p\tau(\max(\boldsymbol{v}) - u\tau)^{p-1}$ and for $u \leq \alpha$, $\hat{\text{RG}}_p^{(\boldsymbol{v})}(u) = \frac{\text{RG}_p(\boldsymbol{v}) - (\max(\boldsymbol{v}) - \alpha\tau)^p}{\alpha}$.

Figure 3 (top) illustrates $\underline{\text{RG}}_p^{(\boldsymbol{v})}$ and the corresponding lower hull $H^{(\boldsymbol{v})}_{\text{RG}_p}$ for example vectors with $p \in \{0.5, 1, 2\}$.

From $\hat{\text{RG}}_p^{(\boldsymbol{v})}$, we can compute for any vector $\boldsymbol{v}$, the minimum possible variance attainable for it by an unbiased nonnegative estimator:

$$\text{VAR}[\hat{\text{RG}}_p^{(\boldsymbol{v})}] = \int_0^1 \hat{\text{RG}}_p^{(\boldsymbol{v})}(u)^2 du - \text{RG}_p(\boldsymbol{v})^2 . \tag{8}$$

We use the $\boldsymbol{v}$-optimal estimates as a reference point to measure the "variance competitiveness" of estimators.

The L estimator.

Theorem 5.2. *[13] The estimator $\hat{\text{RG}}_p^{(L)}$ ($p > 0$), specified as the solution of*

$$\forall \boldsymbol{v} \forall \rho \ \hat{\text{RG}}_p^{(L)}(\boldsymbol{v}, \rho) = -\int_\rho^1 \frac{\frac{d\underline{\text{RG}}_p^{(\boldsymbol{v})}(u)}{du}}{u} du \tag{9}$$

has the following properties:

• *It is nonnegative and unbiased.*

• *It is the unique (up to equivalence) variance$^+$-optimal monotone estimator.*

• *It is $\prec^+$-optimal with respect to the partial order $\prec$*

$$\boldsymbol{v} \prec \boldsymbol{z} \iff \text{RG}_p(\boldsymbol{v}) < \text{RG}_p(\boldsymbol{z}) .$$

• *It has finite variances and is 4-competitive:*

$$\forall \boldsymbol{v}, \text{VAR}[\hat{\text{RG}}_p^{(L)}|\boldsymbol{v}] + \text{RG}_p(\boldsymbol{v})^2 \leq 4(\text{VAR}[\hat{\text{RG}}_p^{(\boldsymbol{v})}] + \text{RG}_p(\boldsymbol{v})^2) .$$
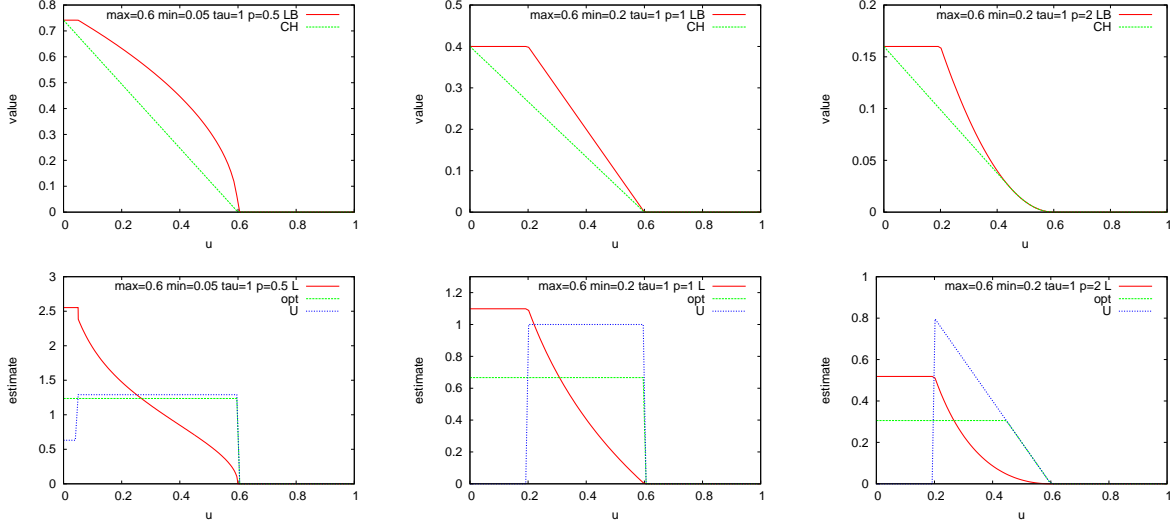
9

Figure 3: Top: The lower bound function and corresponding lower hull for example vectors and $p \in \{0.5, 1, 2\}$. Bottom: the corresponding optimal, L, and U estimates on outcomes consistent with the vector.

$\prec^+$-optimality with respect to this particular order means that any estimator with a strictly lower variance for a data vector must have strictly higher variance on some vector with a smaller range – this means that the L estimator "prioritizes" data where the range (or difference when aggregated) is small. "Competitiveness," is a strong property that means that for *all* data vectors, the variance under the L estimator is not too far off the minimum possible variance for that vector by a nonnegative unbiased estimator.

We solve the equations to derive the L estimator under PPS sampling. For an outcome $S(u, \boldsymbol{v})$, we define $v_{\min} = \min(\boldsymbol{v})$ if $|S| = r$ and $v_{\min} = u\tau$ otherwise.

$$
\begin{array}{c|cc}
 & \multicolumn{2}{c}{\hat{\text{RG}}_p^{(L)}} \\
\hline
|S| = 0 & 0 \\
|S| \geq 1 & (\max(\boldsymbol{v}) - v_{\min})^p \max\{1, \frac{\tau}{v_{\min}}\} - \int_{\min\{1, \frac{v_{\min}}{\tau}\}}^{\min\{1, \frac{\max(\boldsymbol{v})}{\tau}\}} \frac{(\max(\boldsymbol{v}) - x\tau)^p}{x^2} dx
\end{array}
\tag{10}
$$

Estimators and variance for RG and RG_2 are provided in Tables 6 and 7. We obtain the following tight ratios on competitiveness:

$$
\frac{\text{VAR}[\hat{\text{RG}}^{(L)} | \boldsymbol{v}] + \text{RG}(\boldsymbol{v})^2}{\text{VAR}[\hat{\text{RG}}^{(\boldsymbol{v})}] + \text{RG}(\boldsymbol{v})^2} \leq 2
\tag{11}
$$

$$
\frac{\text{VAR}[\hat{\text{RG}_2}^{(L)} | \boldsymbol{v}] + \text{RG}_2(\boldsymbol{v})^2}{\text{VAR}[\hat{\text{RG}_2}^{(\boldsymbol{v})}] + \text{RG}_2(\boldsymbol{v})^2} \leq 2.5
\tag{12}
$$

The U estimator.

Theorem 5.3. *[13] The estimator* $\hat{\text{RG}}_p^{(U)}$ *($p > 0$), specified as the solution of*

$$
\forall S \equiv S(\rho, \boldsymbol{v}),\ \text{R}\hat{\text{G}}_p(\rho, \boldsymbol{v}) = \sup_{\boldsymbol{z} \in V^*(S)} \inf_{0 \leq \eta < \rho} \frac{\text{RG}_p(\eta, \boldsymbol{z}) - \int_\rho^1 \text{R}\hat{\text{G}}_p(u, \boldsymbol{v}) du}{\rho - \eta}
$$

has the following properties:

• *It is nonnegative and unbiased.*

10

● *It is $\prec^+$-optimal with respect to the partial order $\prec$*

$$\boldsymbol{v} \prec \boldsymbol{z} \iff \mathrm{RG}_p(\boldsymbol{v}) > \mathrm{RG}_p(\boldsymbol{z}) .$$

● *It has finite variances for all data vectors.*

The U estimator "prioritizes" data where the range (or difference when aggregated) is large. In particular, it is the nonnegative unbiased estimator with minimum variance on data with $\min(\boldsymbol{v}) = 0$.

The solution for PPS with all-entries equal τ is provided as Algorithm 1 (see Appendix A for calculation).

The estimator $\hat{\mathrm{RG}}_p^{(U)}$ and its variance for $p = 1, 2$ are provided in Tables 8 and 9 (See Appendix B for details).
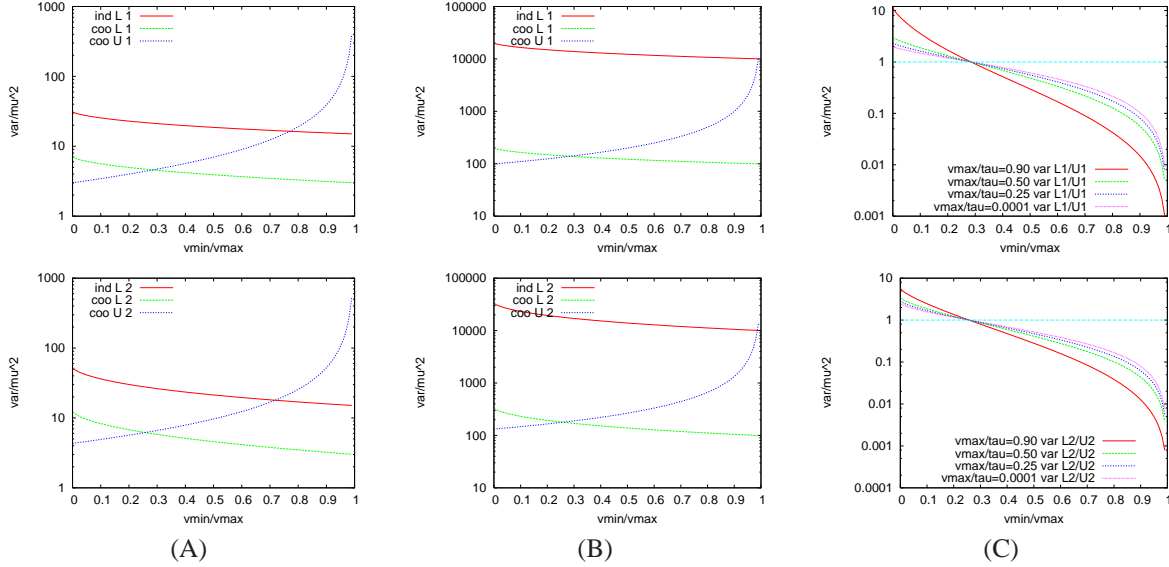

Figure 4: Variance (normalized by square of expectation) of $\hat{\mathrm{RG}}_p^{(L)}$ estimator over independent samples and of $\hat{\mathrm{RG}}_p^{(L)}$ and $\hat{\mathrm{RG}}_p^{(U)}$ over shared-seed samples. Sampling with all-entries equal τ. (A): data with $\max(\boldsymbol{v}) = 0.25\tau$. (B): data with $\max(\boldsymbol{v}) = 0.01\tau$. (C): ratio $\mathrm{VAR}[\hat{\mathrm{RG}}_p^{(L)}]/\mathrm{VAR}[\hat{\mathrm{RG}}_p^{(U)}]$ for shared-seed sampling, selected ratios $\max(\boldsymbol{v})/\tau$. Sweeping $\min(\boldsymbol{v})$. Top shows $p = 1$, bottom is $p = 2$.

Choosing an estimator. How to choose between the L and U estimators? Figure 3 shows the $\boldsymbol{v}$-optimal estimates and the L and U estimators for example vectors, illustrating their form and monotonicity of L and $\boldsymbol{v}$-optimal.

The estimators and their variance depends only on τ and the maximum and minimum entry values $\max(\boldsymbol{v})$ and $\min(\boldsymbol{v})$. For all-entries-equal τ and $\max(\boldsymbol{v}) \leq \tau$, we study the variance dependence on the ratio $\frac{\min(\boldsymbol{v})}{\max(\boldsymbol{v})}$. The variance is 0 when $\mathrm{RG}(\boldsymbol{v}) = 0$. The estimator $\hat{\mathrm{RG}}_p^{(U)}$ has lower variance when $\frac{\min(\boldsymbol{v})}{\max(\boldsymbol{v})}$ is sufficiently small. The solution ϕ_p of $\mathrm{VAR}[\hat{\mathrm{RG}}_p^{(U)}|\boldsymbol{v}] = \mathrm{VAR}[\hat{\mathrm{RG}}_p^{(L)}|\boldsymbol{v}]$ for $x = \frac{\min(\boldsymbol{v})}{\max(\boldsymbol{v})}$, is such that

$$\mathrm{VAR}[\hat{\mathrm{RG}}_p^{(U)}|\boldsymbol{v}] < \mathrm{VAR}[\hat{\mathrm{RG}}_p^{(L)}|\boldsymbol{v}] \iff \frac{\min(\boldsymbol{v})}{\max(\boldsymbol{v})} < \phi_p .$$

For $p = 1$, $\phi_1 \approx 0.285$ (is the solution of the equality $(1 - x)/(2x) = \ln(1/x)$). For $p = 2$, $\phi_2 \approx 0.258$.

This suggests selecting an estimator according to expected characteristics of the data. If we expect $\mathrm{RG}(\boldsymbol{v}) \geq (1 - \phi_p) \max(\mathbf{v})$, we choose $\hat{\mathrm{RG}}_p^{(U)}$ and otherwise choose $\hat{\mathrm{RG}}_p^{(L)}$.

The variance of the L estimator over independent samples and of the L and U estimators over shared-seed samples is illustrated in Figure 4. The figure also illustrates the relation between the variance of the shared-seed L and U

dataset	# keys	p1%	p2%	$\sum_{i,h} v_i(h)$	p1%	p2%	$L_1/\sum_{i,h} v_i(h)$	$L_{1+}/\sum_{i,h} v_i(h)$	$L_{1-}/\sum_{i,h} v_i(h)$
destIP	3.8×10^4	65%	65%	1.1×10^6	49%	51%	0.36	0.19	0.18
Server	2.7×10^5	53%	56%	2.9×10^6	50%	50%	0.75	0.38	0.37
Surnames	1.9×10^4	100%	100%	8.9×10^7	48.6%	51.4%	0.094	0.0617	0.0327
OSPD8	7.5×10^5	99%	99%	1.57×10^{10}	46.8%	53.2%	0.0826	0.0727	0.0099

Table 5: Datasets. Table shows total number of distinct keys with positive value in at least one of the two instances, corresponding percentage in each instance, total sum of values $\sum_{i,h} v_i(h) \equiv \sum_{h \in K} \sum_{i \in [2]} v_i(h)$, and fraction (shown as percentage) in instance $i = 1, 2$: $\frac{\sum_{h \in K} v_i(h)}{\sum_{i,h} v_i(h)}$, and normalized L_1, L_{1+} and L_{1-} differences.

estimators. When $\max(\boldsymbol{v})/\tau \ll 1$ (which we expect to be a prevailing scenario), $\text{VAR}[\hat{\text{RG}}^{(L)}]$ is nearly at most 2 times $\text{VAR}[\hat{\text{RG}}^{(U)}]$ but as $\min(\boldsymbol{v}) \to \max(\boldsymbol{v})$, the ratio $\text{VAR}[\hat{\text{RG}}^{(U)}]/\text{VAR}[\hat{\text{RG}}^{(L)}]$ is not bounded. When $\max(\boldsymbol{v})/\tau$ is close to 1, the variance of the U estimator is close to 0, and $\text{VAR}[\hat{\text{RG}}^{(L)}]/\text{VAR}[\hat{\text{RG}}^{(U)}]$ is not bounded. Interestingly, for $p = 2$, the variance of the U estimator is always at least $\frac{4}{3}\text{RG}_4(\boldsymbol{v})$, and thus, using (12), the variance of the L estimator is at most 4.4 times the variance of the $\boldsymbol{v}$-optimal (and thus of the U estimator).

6 Experimental Evaluation

We study the estimation quality of our L_p^p estimators, recalling that our L_p^p estimate is the sum of RG_p estimates: $\hat{L}_p^p = \sum_{h \in K} \hat{\text{RG}}_p(h)$. This estimator is unbiased and has expectation $L_p^p = \sum_{h \in K} \text{RG}_p(\boldsymbol{v}(h))$. The variance is additive and is $\sum_{h \in K} \text{VAR}[\hat{\text{RG}}_p(h)]$. The squared coefficient of variation is the ratio of the variance and the square of the expectation: $\text{CV}^2(\hat{L}_p^p) = \frac{\sum_{h \in K} \text{VAR}[\hat{\text{RG}}_p|\boldsymbol{v}(h)]}{(\sum_{h \in K} \text{RG}_p(\boldsymbol{v}(h)))^2}$.

Figure 5 shows the squared coefficient of variation CV^2 of our L_p^p estimators as a function of the sampled fraction of the dataset. Each of the two instances was subjected to Poisson PPS [22] sampling (Results are essentially identical for Priority sampling [27, 18]). We study accuracy when applying the single-key estimator $\hat{\text{RG}}_p^{(L)}$ for independent samples of instances and the estimators $\hat{\text{RG}}_p^{(L)}$ and $\hat{\text{RG}}_p^{(U)}$ for shared-seed (coordinated) samples of the two instances, for $p = 1, 2$.

We used 4 datasets with properties summarized in Table 5.

destIP: keys: (anonymized) IP destination addresses. value: the number of IP flows to this destination IP (source: IP packet traces). Instances: two consecutive time periods.

Server: Keys: (anonymized) source IP addresses. values: the number of HTTP requests issued to the server from this address. Instances: two consecutive time periods. Source: Web server log.

Surnames: Keys: the 18.5×10^3 most common surnames in the US. value: the number of occurrences of the surname in English books digitized by Google and published within the time period [26]. Instances: the years 2007 and 2008.

OSPD8: Keys: 7.5×10^4 8 letter words that appear in the Official Scrabble Players Dictionary (OSPD). value: the number of occurrences of the term in English books digitized by Google and published within a time period [26]. Instances: the years 2007 and 2008.

We can see qualitatively, that all estimators, even over independent samples, are satisfactory, in that the CV is small for a sample that is a small fraction of the full data set. The monotone estimator $\hat{\text{RG}}_p^{(L)}$ over coordinated (shared-seed) samples outperforms, by orders of magnitude, the monotone estimator $\hat{\text{RG}}_p^{(L)}$ over independent samples. The gap widens for more aggressive sampling.

The first two datasets (destIP and Server) exhibit significant difference between instances: the L_1 distance is a large fraction of the total sum of values $\sum_{h \in K} \sum_{i \in [2]} v_i(h)$. On these datasets, $\hat{\text{RG}}_p^{(U)}$ outperforms $\hat{\text{RG}}_p^{(L)}$ on shared-seed samples. The last two datasets (Surnames and OSPD8) have small difference between instances and $\hat{\text{RG}}_p^{(L)}$ outperforms $\hat{\text{RG}}_p^{(U)}$ on shared seed samples. These trends are more pronounced for the higher moment $p = 2$. In this case, on Surnames and OSPD8 datasets, $\hat{\text{RG}}_p^{(L)}$ over independent samples outperform $\hat{\text{RG}}_p^{(U)}$ over shared-

seed samples. We can see that we can significantly improve accuracy by tailoring the selection of the estimator to properties of the data. The performance of the U estimator, however, can significantly diverge with similarity whereas the competitive L estimator is guaranteed not to be too far off. Therefore, when there is no prior knowledge on the difference, we suggest using the L estimator.

The datasets also differ in the symmetry of change. The change is more symmetric in the first two data sets $L_{p+} \approx L_{p-}$ whereas there is a general growth trend $L_{p+} \gg L_{p-}$ in the last two datasets. We did not include performance figures for the asymmetric differences $\sum_{h \in K} \text{RG}_{p+}(\boldsymbol{v}(h))$ and $\sum_{h \in K} \text{RG}_{p-}(\boldsymbol{v}(h))$, but trends are similar to the symmetric variants.

7 Conclusion

Difference queries are essential for monitoring, planning, and anomaly and change detection. Random sampling is an important tool for retaining the ability to query data under resource limitations. We provide the first satisfactory solution for estimating differences over sampled data sets. Our solution is comprehensive, covering common sampling schemes. It is supported by rigorous analysis and novel techniques that also allow us to gain deeper understanding and establish optimality. We demonstrated that our estimators perform well on diverse data sets.

References

[1] K. S. Beyer, P. J. Haas, B. Reinwald, Y. Sismanis, and R. Gemulla. On synopses for distinct-value estimation under multiset operations. In *SIGMOD*, pages 199–210. ACM, 2007.

[2] K. R. W. Brewer, L. J. Early, and S. F. Joyce. Selecting several samples from a single population. *Australian Journal of Statistics*, 14(3):231–239, 1972.

[3] A. Z. Broder. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences*, pages 21–29. IEEE, 1997.

[4] A. Z. Broder. Identifying and filtering near-duplicate documents. In *Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching*, volume 1848 of *LLNCS*, pages 1–10. Springer, 2000.

[5] M. Charikar, S. Chaudhuri, R. Motwani, and V. Narasayya. Towards estimation error guarantees for distinct values. In *Proceedings of ACM Principles of Database Systems*, 2000.

[6] S. S. Chawathe, A. Rajaraman, H. Garcia-Molina, and J. Widom. Change detection in hierarchically structured information. In *Proceedings of the 1996 ACM SIGMOD international conference on Management of data*, 1996.

[7] E. Cohen. Size-estimation framework with applications to transitive closure and reachability. *J. Comput. System Sci.*, 55:441–453, 1997.

[8] E. Cohen, N. Duffield, C. Lund, M. Thorup, and H. Kaplan. Efficient stream sampling for variance-optimal estimation of subset sums. *SIAM J. Comput.*, 40(5), 2011.

[9] E. Cohen and H. Kaplan. Summarizing data using bottom-k sketches. In *Proceedings of the ACM PODC'07 Conference*, 2007.

[10] E. Cohen and H. Kaplan. Tighter estimation using bottom-k sketches. In *Proceedings of the 34th VLDB Conference*, 2008.

[11] E. Cohen and H. Kaplan. Leveraging discarded samples for tighter estimation of multiple-set aggregates. In *Proceedings of the ACM SIGMETRICS'09 Conference*, 2009.

[12] E. Cohen and H. Kaplan. Get the most out of your sample: Optimal unbiased estimators using partial information. In *Proc. of the 2011 ACM Symp. on Principles of Database Systems (PODS 2011)*. ACM, 2011. full version: http://arxiv.org/abs/1203.4903.
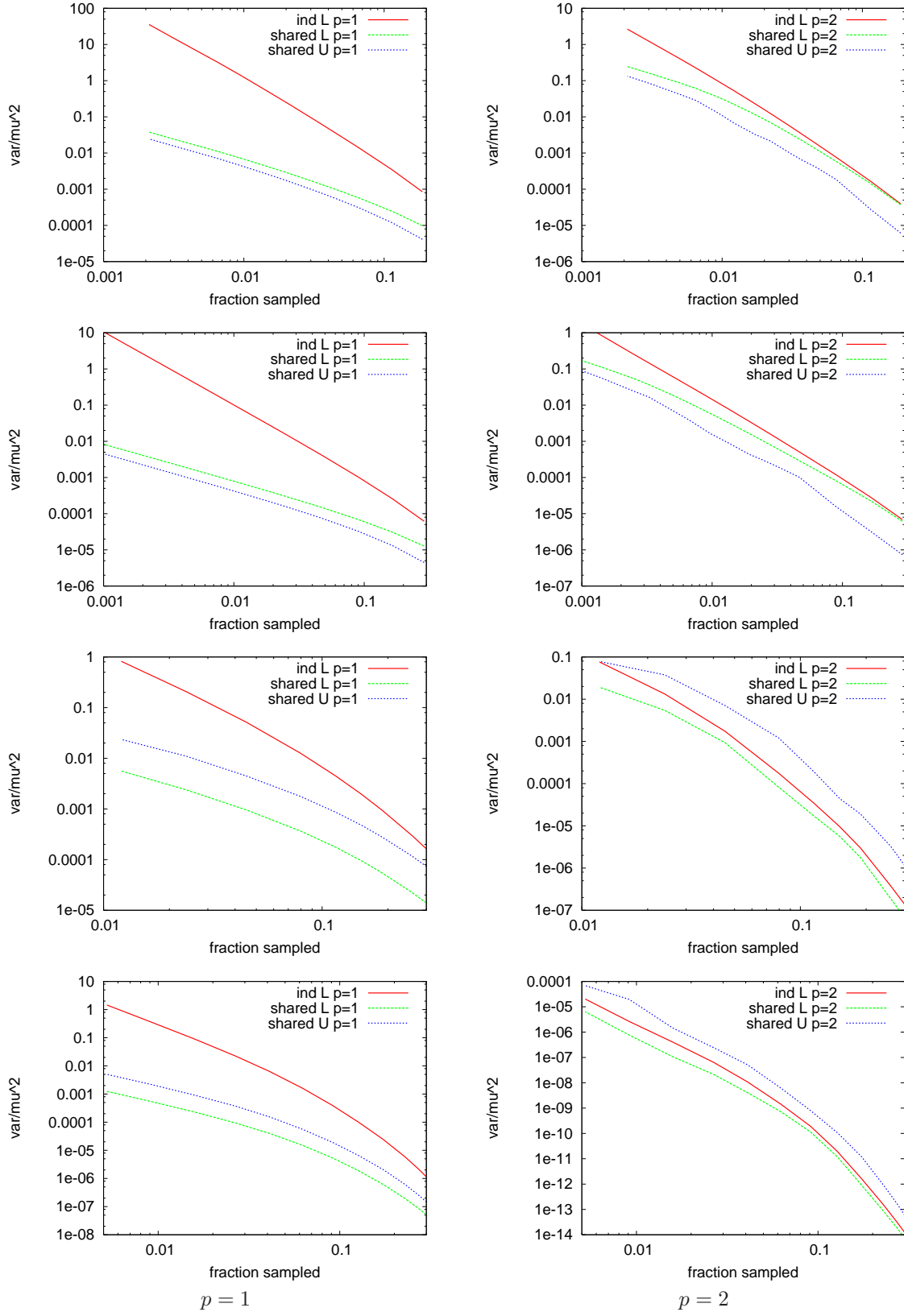
Figure 5: Datasets top to bottom: destIP, Server, Surnames, OSPD8. CV^2 for fraction of sampled items.

[13] E. Cohen and H. Kaplan. A case for customizing estimators: Coordinated samples. Technical Report cs.ST/1212.0243, arXiv, 2012.

[14] E. Cohen and H. Kaplan. What you can do with coordinated samples. Technical Report cs.DB/1206.5637, arXiv, 2012.

[15] E. Cohen, H. Kaplan, and S. Sen. Coordinated weighted sampling for estimating aggregates over multiple weight assignments. *Proceedings of the VLDB Endowment*, 2(1–2), 2009. full version: `http://arxiv.org/abs/0906.4560`.

[16] T. Dasu, T. Johnson, S. Muthukrishnan, and V. Shkapenyuk. Mining database structure; or, how to build a data quality browser. In *Proc. SIGMOD Conference*, pages 240–251, 2002.

[17] W. Dong, M. Charikar, and K. Li. Asymmetric distance estimation with sketches for similarity search in high-dimensional spaces. In *SIGIR*, 2008.

[18] N. Duffield, M. Thorup, and C. Lund. Priority sampling for estimating arbitrary subset sums. *J. Assoc. Comput. Mach.*, 54(6), 2007.

[19] P. Gibbons and S. Tirthapura. Estimating simple functions on the union of data streams. In *Proceedings of the 13th Annual ACM Symposium on Parallel Algorithms and Architectures*. ACM, 2001.

[20] P. B. Gibbons. Distinct sampling for highly-accurate answers to distinct values queries and event reports. In *International Conference on Very Large Databases (VLDB)*, pages 541–550, 2001.

[21] M. Hadjieleftheriou, X. Yu, N. Koudas, and D. Srivastava. Hashed samples: Selectivity estimators for set similarity selection queries. In *Proceedings of the 34th VLDB Conference*, 2008.

[22] J. Hájek. *Sampling from a finite population*. Marcel Dekker, New York, 1981.

[23] D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

[24] D. Kifer, S. Ben-David, and J. Gehrke. Detecting change in data streams. In *VLDB*, 2004.

[25] D.E. Knuth. *The Art of Computer Programming, Vol. 2: Seminumerical Algorithms*. Addison-Wesley, 1969.

[26] J. B. Michel, Y. K. Shen, A. Presser Aiden, A. Veres, M. K. Gray, W. Brockman, The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011. `http://www.sciencemag.org/content/331/6014/176`.

[27] E. Ohlsson. Sequential poisson sampling. *J. Official Statistics*, 14(2):149–162, 1998.

[28] E. Ohlsson. Coordination of pps samples over time. In *The 2nd International Conference on Establishment Surveys*, pages 255–264. American Statistical Association, 2000.

[29] B. Rosén. Asymptotic theory for successive sampling with varying probabilities without replacement, I. *The Annals of Mathematical Statistics*, 43(2):373–397, 1972.

[30] B. Rosén. Asymptotic theory for order sampling. *J. Statistical Planning and Inference*, 62(2):135–158, 1997.

[31] P. J. Saavedra. Fixed sample size pps approximations with a permanent random number. In *Proc. of the Section on Survey Research Methods, Alexandria VA*, pages 697–700. American Statistical Association, 1995.

[32] M. Szegedy. The DLT priority sampling is essentially optimal. In *Proc. 38th Annual ACM Symposium on Theory of Computing*. ACM, 2006.

[33] J.S. Vitter. Random sampling with a reservoir. *ACM Trans. Math. Softw.*, 11(1):37–57, 1985.

Appendix

$$
\begin{array}{ll}
\hline
& \hat{\text{RG}}^{(L)} \\
\hline
|S| = 0 & 0 \\
|S| \geq 1 & \max\{\max(\boldsymbol{v}) - \tau, 0\} - \max\{\min(\boldsymbol{v}) - \tau, 0\} + \tau \ln \frac{\min\{\max(\boldsymbol{v}), \tau\}}{\min\{v_{\min}, \tau\}} \\
\hline
\end{array}
\tag{13}
$$

$$
\begin{array}{ll}
\hline
\text{Condition} & \text{VAR}[\hat{\text{RG}}^{(L)} | \boldsymbol{v}] \\
\hline
\min(\boldsymbol{v}) \geq \tau & 0 \\
\max(\boldsymbol{v}) \leq \tau,\ \min(\boldsymbol{v}) = 0 & 2\text{RG}(\boldsymbol{v})\tau - \text{RG}(\boldsymbol{v})^2 \\
\max(\boldsymbol{v}) \leq \tau,\ \min(\boldsymbol{v}) > 0 & 2\text{RG}(\boldsymbol{v})\tau - \text{RG}(\boldsymbol{v})^2 - 2\tau \min(\boldsymbol{v}) \ln(\frac{\max(\boldsymbol{v})}{\min(\boldsymbol{v})}) \\
0 < \min(\boldsymbol{v}) \leq \tau \leq \max(\boldsymbol{v}) & (\tau)^2 - \min(\boldsymbol{v})^2 - 2\tau \min(\boldsymbol{v}) \ln(\frac{\tau}{\min(\boldsymbol{v})}) \\
0 = \min(\boldsymbol{v}),\ \tau \leq \max(\boldsymbol{v}) & (\tau)^2 - \min(\boldsymbol{v})^2 \\
\hline
\end{array}
\tag{14}
$$

Table 6: The estimator $\hat{\text{RG}}^{(L)}$ (top) and variance for data $\boldsymbol{v}$ (bottom) for shared-seed sampling

$$
\begin{array}{ll}
\hline
& \hat{\text{RG}}_2^{(L)} \\
\hline
|S| = 0 & 0 \\
|S| \geq 1 & \max\{\max(\boldsymbol{v}), \tau\}^2 - \max\{\min(\boldsymbol{v}), \tau\}^2 - 2\max\{\min(\boldsymbol{v}), \tau\}(\max(\boldsymbol{v}) - v_{\min}) \\
& \quad + 2\tau \max(\boldsymbol{v}) \ln \frac{\min\{\max(\boldsymbol{v}), \tau\}}{\min\{v_{\min}, \tau\}} \\
\hline
\end{array}
\tag{15}
$$

$$
\begin{array}{ll}
\hline
\text{Condition} & \text{VAR}[\hat{\text{RG}}_2^{(L)} | \boldsymbol{v}] \\
\hline
\min(\boldsymbol{v}) \geq \tau & 0 \\
\hline
\max(\boldsymbol{v}) \leq \tau & -4\tau \max(\boldsymbol{v}) \min(\boldsymbol{v}) \ln(\frac{\max(\boldsymbol{v})}{\min(\boldsymbol{v})})(2\max(\boldsymbol{v}) - \min(\boldsymbol{v})) - (\max(\boldsymbol{v}) - \min(\boldsymbol{v}))^4 \\
& \quad + \frac{2\tau}{3}(5\max(\boldsymbol{v})^3 + 4\min(\boldsymbol{v})^3 - 9\max(\boldsymbol{v})\min(\boldsymbol{v})^2) \\
\hline
\min(\boldsymbol{v}) \leq \tau & 4\max(\boldsymbol{v})\min(\boldsymbol{v})\tau(\min(\boldsymbol{v}) - 2\max(\boldsymbol{v})) \ln \frac{\tau}{\min(\boldsymbol{v})} \\
\wedge & \quad + 4\max(\boldsymbol{v})\min(\boldsymbol{v})(\tau)^2 + \frac{(\tau)^4}{3} + \frac{8\min(\boldsymbol{v})^3\tau}{3} \\
\max(\boldsymbol{v}) \geq \tau & -6\max(\boldsymbol{v})\min(\boldsymbol{v})^2\tau - 4\max(\boldsymbol{v})^2\min(\boldsymbol{v})^2 \\
& \quad -\min(\boldsymbol{v})^4 + 4\max(\boldsymbol{v})\min(\boldsymbol{v})^3 + 4\max(\boldsymbol{v})^2(\tau)^2 - 2\max(\boldsymbol{v})(\tau)^3 \\
\hline
\end{array}
\tag{16}
$$

Table 7: The estimator $\hat{\text{RG}}_2^{(L)}$ (top) and variance for data $\boldsymbol{v}$ (bottom) for shared-seed sampling

A   Derivation of $\hat{\text{RG}}_p^{(U)}$

If $\rho\tau > \max(\boldsymbol{v})$ then $\underline{\text{RG}}_p(\rho, \boldsymbol{v}) = 0$ and $\hat{\text{RG}}_p^{(U)}(\rho, \boldsymbol{v}) = 0$. Otherwise, when $\min(\boldsymbol{v}) < \rho\tau \leq \max(\boldsymbol{v})$, noting that the supremum is obtained by a vector $\boldsymbol{v}'$ with maximum entry $\max(\boldsymbol{v})$ and minimum entry 0,

$$
\hat{\text{RG}}_p^{(U)}(\rho, \boldsymbol{v}) = \inf_{0 \leq \eta < \rho} \frac{(\max(\boldsymbol{v}) - \eta\tau)^p - \int_\rho^{\min\{1, \frac{\max(\boldsymbol{v})}{\tau}\}} \hat{\text{RG}}_p^{(U)}(u, \boldsymbol{v}) du}{\rho - \eta}
\tag{17}
$$

condition	$\hat{\mathrm{RG}}^{(U)}$
$|S|=0$	0
$1\le|S|\le r-1$	$\max\{\tau,\max(\boldsymbol{v})\}$
$|S|=r$	$\max\{\max(\boldsymbol{v}),\tau\}-\max\{\min(\boldsymbol{v}),\tau\}$

condition on $\boldsymbol{v}$	$\mathrm{VAR}[\hat{\mathrm{RG}}^{(U)}\mid\boldsymbol{v}]$
$\min(\boldsymbol{v})\ge\tau$	0
$\max(\boldsymbol{v})\le\tau$	$\mathrm{RG}(\boldsymbol{v})(\tau-\mathrm{RG}(\boldsymbol{v}))$
$\min(\boldsymbol{v})<\tau<\max(\boldsymbol{v})$	$\min(\boldsymbol{v})(\tau-\min(\boldsymbol{v}))$

Table 8: The estimator $\hat{\mathrm{RG}}^{(U)}$ (left) and its variance $\mathrm{VAR}[\hat{\mathrm{RG}}^{(U)}\mid\boldsymbol{v}]$ (right) for shared-seed sampling.

condition on $S(\rho,\boldsymbol{v})$	$\hat{\mathrm{RG}}_2^{(U)}(S)$
$\frac{\max(\boldsymbol{v})}{\tau}\ge 2,\ \rho>\frac{\min(\boldsymbol{v})}{\tau}$	$\max(\boldsymbol{v})^2$
$\frac{\max(\boldsymbol{v})}{\tau}\ge 2,\ \rho\le\frac{\min(\boldsymbol{v})}{\tau}$	$\max(\boldsymbol{v})^2-2\tau\max(\boldsymbol{v})+\min(\boldsymbol{v})\tau$
$\frac{\max(\boldsymbol{v})}{\tau}\le 1,\ \rho\in(\frac{\min(\boldsymbol{v})}{\tau},\frac{\max(\boldsymbol{v})}{\tau}]$	$2\tau(\max(\boldsymbol{v})-\rho\tau)$
$\frac{\max(\boldsymbol{v})}{\tau}\le 1,\ \rho\le\frac{\min(\boldsymbol{v})}{\tau}$	0
$\frac{\max(\boldsymbol{v})}{\tau}\le 1,\ \rho\ge\frac{\max(\boldsymbol{v})}{\tau}$	0
$\frac{\max(\boldsymbol{v})}{\tau}\in[1,2],\ \rho>2-\frac{\max(\boldsymbol{v})}{\tau},\ \rho>\frac{\min(\boldsymbol{v})}{\tau}$	$4\tau(\max(\boldsymbol{v})-\tau)$
$\frac{\max(\boldsymbol{v})}{\tau}\in[1,2],\ \rho<2-\frac{\max(\boldsymbol{v})}{\tau},\ \rho>\frac{\min(\boldsymbol{v})}{\tau}$	$2\tau(\max(\boldsymbol{v})-\rho\tau)$
$\frac{\max(\boldsymbol{v})}{\tau}\in[1,2],\ \rho<2-\frac{\max(\boldsymbol{v})}{\tau},\ \rho<\frac{\min(\boldsymbol{v})}{\tau}$	0
$\frac{\max(\boldsymbol{v})}{\tau}\in[1,2],\ \rho\le\frac{\min(\boldsymbol{v})}{\tau}>2-\frac{\max(\boldsymbol{v})}{\tau}$	$\frac{\tau}{\min(\boldsymbol{v})}\mathrm{RG}_2(\boldsymbol{v})-4\tau(\max(\boldsymbol{v})-\tau)(\frac{\tau}{\min(\boldsymbol{v})}-1)$

condition on $\boldsymbol{v}$	$\mathrm{VAR}[\hat{\mathrm{RG}}_2^{(U)}\mid\boldsymbol{v}]$
$\min(\boldsymbol{v})\ge\tau$	0
$\max(\boldsymbol{v})\le\tau$	$\mathrm{RG}_3(\boldsymbol{v})(\frac{4}{3}\tau-\mathrm{RG}(\boldsymbol{v}))$
$\frac{\max(\boldsymbol{v})}{\tau}\in[1,2],\ \frac{\min(\boldsymbol{v})}{\tau}\ge 2-\frac{\max(\boldsymbol{v})}{\tau}$	$\frac{2(\tau-\min(\boldsymbol{v}))}{\tau}\left(\mathrm{RG}_2(\boldsymbol{v})\right)-4\tau(\max(\boldsymbol{v})-\tau)\Big)^2$
$\frac{\max(\boldsymbol{v})}{\tau}\in[1,2],\ \frac{\min(\boldsymbol{v})}{\tau}<2-\frac{\max(\boldsymbol{v})}{\tau}$	$\frac{\max(\boldsymbol{v})-\tau}{\tau}\Big(4\tau(\max(\boldsymbol{v})-\tau)-\mathrm{RG}_2(\boldsymbol{v})\Big)^2+\frac{\min(\boldsymbol{v})}{\tau}\mathrm{RG}_4(\boldsymbol{v})$ $+\frac{(\mathrm{RG}_2(\boldsymbol{v})+4(\tau)^2-4\max(\boldsymbol{v})\tau)^3-\mathrm{RG}_3(\boldsymbol{v})(\mathrm{RG}(\boldsymbol{v})-2\tau)^3}{6(\tau)^2}$
$\frac{\max(\boldsymbol{v})}{\tau}\ge 2$	$(2\max(\boldsymbol{v})-\min(\boldsymbol{v}))^2\min(\boldsymbol{v})(\tau-\min(\boldsymbol{v}))$

Table 9: The estimator $\hat{\mathrm{RG}}_2^{(U)}$ (top) and its variance $\mathrm{VAR}[\hat{\mathrm{RG}}_2^{(U)}\mid\boldsymbol{v}]$ (bottom) for shared-seed sampling.

Algorithm 1 $\hat{\text{RG}}_p^{(U)}(S)$

---

   if $|S| = 0$ **then return** 0 $\triangleright$ from hereafter $|S| > 0$

   $m \leftarrow \max_{i \in S} v_i$ $\triangleright$ $m = \max(\boldsymbol{v})$

   if $|S| < r$ **then** $n \leftarrow 0$

   else $n \leftarrow \min_{i \in S} v_i$ $\triangleright$ $n = \underline{\min}(S)$

   if $n \geq \tau$ **then return** $(m - n)^p$ $\triangleright$ case: $\min(\boldsymbol{v}) \geq \tau$

   if $p \leq 1$ **then** $\triangleright$ case: $\min(\boldsymbol{v}) \leq \tau$ and $p \leq 1$

      if n=0 **then return** $m^p \frac{\tau}{\min\{m,\tau\}}$

      else   **return** $\frac{\tau}{n}\left((m-n)^p - \frac{\min\{m,\tau\}-n}{\min\{m,\tau\}} m^p\right)$

   if $m \leq \tau$ **then** $\triangleright$ case: $\max(\boldsymbol{v}) \leq \tau, p > 1$

      if $\rho\tau > n$ **then return** $p\tau(m - \rho\tau)^{p-1}$

      else   **return** 0

                                     $\triangleright$ case: $n < \tau < \max(\boldsymbol{v})$ and $p > 1$

   $\eta_0 \leftarrow \frac{p\tau - m}{(p-1)\tau}$

   if $\eta_0 \in (0, 1)$ **then** $\triangleright$ subcase: $\eta_0 \in (0, 1)$

      if $\rho \geq \max\{\eta_0, n/\tau\}$ **then return** $\frac{(m-\eta_0\tau)^p}{1-\eta_0}$

      if $n/\tau < \rho < \eta_0$ **then return** $p\tau(m - \rho\tau)^{p-1}$

      if $\rho \leq n/\tau \leq \eta_0$ **then return** 0

      if $\rho \leq n/\tau \geq \eta_0$ **then**

         return $\frac{\tau(m-n)^p}{n} - \frac{(\tau-n)(m-\eta_0\tau)^p}{n(1-\eta_0)}$

   else $\triangleright$ subcase: $\eta_0 \notin (0, 1)$

      if $\rho\tau > n$ **then return** m^p

      else   **return** $\frac{\tau}{n}(m - n)^p - m^p\left(\frac{\tau}{n} - 1\right)$

If $\rho\tau \leq \min(\boldsymbol{v})$,

$$
\begin{aligned}
\hat{\text{RG}}_p^{(U)}(\rho, \boldsymbol{v}) &= \frac{\text{RG}_p(\boldsymbol{v}) - \int_{\rho}^{\min\{1, \frac{\max(\boldsymbol{v})}{\tau}\}} \hat{\text{RG}}_p^{(U)}(u, \boldsymbol{v})du}{\rho} \\
&= \frac{\text{RG}_p(\boldsymbol{v}) - \int_{\min\{1, \frac{\min(\boldsymbol{v})}{\tau}\}}^{\min\{1, \frac{\max(\boldsymbol{v})}{\tau}\}} \hat{\text{RG}}_p^{(U)}(u, \boldsymbol{v})du}{\min\{1, \min(\boldsymbol{v})/\tau\}}
\end{aligned}
\tag{18}
$$

If $\min(\boldsymbol{v}) \geq \tau$, $|S| = r$, $\hat{\text{RG}}_p^{(U)} = \hat{\text{RG}}_p^{(L)} \equiv \text{RG}_p(\boldsymbol{v})$. If $\max(\boldsymbol{v}) \leq \tau$, $\int_{\rho}^{1} \hat{\text{RG}}_p(u, \boldsymbol{v})du = \underline{\text{RG}}_p(\rho, \boldsymbol{v})$ and the infimum is the derivative of the lower bound function, and thus, $\hat{\text{RG}}_p(\rho, \boldsymbol{v}) = p\tau(\max(\boldsymbol{v}) - \rho\tau)^{p-1}$.

$$
\begin{array}{c|c}
|S| & \hat{\text{RG}}_p^{(U)} \\
\hline
0, r : & 0 \\
1 \ldots r-1 : & p\tau(\max(\boldsymbol{v}) - \rho\tau)^{p-1}
\end{array}
\tag{19}
$$

We now consider the case $\min(\boldsymbol{v}) \leq \tau \leq \max(\boldsymbol{v})$, solving (17) for $\rho > \min(\boldsymbol{v})/\tau$. For $\rho = 1$ we obtain the equation

$$
\hat{\text{RG}}_p^{(U)}(1, \boldsymbol{v}) = \inf_{0 \leq \eta < \rho} \frac{(\max(\boldsymbol{v}) - \eta\tau)^p}{1 - \eta} \ .
$$

When $p = 1$, the derivative is positive and the infimum is $\max(\boldsymbol{v})$. We obtain that $\hat{\text{RG}}^{(U)}(\rho, \boldsymbol{v}) = \max(\boldsymbol{v})$ for $\rho \geq \min(\boldsymbol{v})/\tau$. Using (18), $\hat{\text{RG}}^{(U)}(\rho, \boldsymbol{v}) = \max(\boldsymbol{v}) - \tau$ when $\rho \leq \frac{\min(\boldsymbol{v})}{\tau}$.

For $p \neq 1$, we need to find the value where $h(\eta) = \frac{(\max(\boldsymbol{v}) - \eta \tau)^p}{1 - \eta}$ is minimized. The derivative is

$$\frac{\partial h(\eta)}{\partial \eta} = \frac{(\max(\boldsymbol{v}) - \eta \tau)^{p-1}}{1 - \eta}\left(-\tau p + \frac{\max(\boldsymbol{v}) - \eta \tau}{1 - \eta}\right).$$

The derivative is 0 at

$$\eta_0 = \frac{p\tau - \max(\boldsymbol{v})}{\tau(p-1)}.$$

If η_0 is outside $(0,1)$, the infimum is obtained at $\eta = 0$ and the estimate is $\hat{\mathrm{RG}}^{(U)}(\rho, \boldsymbol{v}) = \max(\boldsymbol{v})^p$ for $\rho \geq \min(\boldsymbol{v})/\tau$ and, using (18),

$$\hat{\mathrm{RG}}^{(U)}(\rho, \boldsymbol{v}) = \frac{\tau}{\min(\boldsymbol{v})}\mathrm{RG}_p(\boldsymbol{v}) - \max(\boldsymbol{v})^p\left(\frac{\tau}{\min(\boldsymbol{v})} - 1\right)$$

for $\rho < \min(\boldsymbol{v})/\tau$.

Otherwise, if $\eta_0 \in (0,1)$, the infimum is achieved at η_0. Using (17), the estimate is

$$
\begin{aligned}
\hat{\mathrm{RG}}_p(\rho, \boldsymbol{v}) &= \frac{(\max(\boldsymbol{v}) - \eta_0 \tau)^p}{1 - \eta_0} \\
&= \frac{\tau(p-1)(\max(\boldsymbol{v})^{\frac{p-2}{p-1}} - \tau\frac{p}{p-1})^p}{\max(\boldsymbol{v}) - \tau}
\end{aligned}
$$

for $\rho \in [\max\{\eta_0, \frac{\min(\boldsymbol{v})}{\tau}\}, 1]$ and $\hat{\mathrm{RG}}_p(\rho, \boldsymbol{v}) = p\tau(\max(\boldsymbol{v}) - \rho\tau)^{p-1}$ for $\rho \in (\frac{\min(\boldsymbol{v})}{\tau}, \eta_0)$. Using (18), when $\rho \leq \frac{\min(\boldsymbol{v})}{\tau}$, then $\hat{\mathrm{RG}}_p(\rho, \boldsymbol{v}) = 0$ when $\frac{\min(\boldsymbol{v})}{\tau} < \eta_0$ and

$$\hat{\mathrm{RG}}_p(\rho, \boldsymbol{v}) = \frac{\mathrm{RG}_p(\boldsymbol{v}) - (\frac{\min(\boldsymbol{v})}{\tau} - \eta_0)\frac{(\max(\boldsymbol{v}) - \eta_0 \tau)^p}{1 - \eta_0}}{1 - \frac{\min(\boldsymbol{v})}{\tau}}$$

when $\frac{\min(\boldsymbol{v})}{\tau} \geq \eta_0$.

B Variance of $\hat{\mathrm{RG}}^{(U)}$ and $\hat{\mathrm{RG}}_2^{(U)}$

The estimators $\hat{\mathrm{RG}}^{(U)}$ and $\hat{\mathrm{RG}}_2^{(U)}$, provided in Tables 8 and 9, are obtained by substituting $p = 1$ and $p = 2$ respectively in Algorithm 1. We calculate the variance of these estimators.

Variance of $\hat{\mathrm{RG}}^{(U)}$: When $\max(\boldsymbol{v}) \leq \tau$, we have $\hat{\mathrm{RG}}^{(U)} = \tau$ for $\rho \in (\frac{\min(\boldsymbol{v})}{\tau}, \frac{\max(\boldsymbol{v})}{\tau}]$ and $\hat{\mathrm{RG}}^{(U)} = 0$ otherwise. Hence,

$$\mathrm{VAR}[\hat{\mathrm{RG}}^{(U)}] = (\mathrm{RG}(\boldsymbol{v})^2(1 - \mathrm{RG}(\boldsymbol{v})/\tau) + (\tau - \mathrm{RG}(\boldsymbol{v}))^2\mathrm{RG}(\boldsymbol{v})/\tau = \mathrm{RG}(\boldsymbol{v})\tau - \mathrm{RG}(\boldsymbol{v})^2$$

Variance of $\hat{\mathrm{RG}}_2^{(U)}$: When $\max(\boldsymbol{v}) > \tau$, we have $\eta_0 = 2 - \frac{\max \boldsymbol{v}}{\tau}$. Thus $\eta_0 \in (0, 1) \iff \frac{\max(\boldsymbol{v})}{\tau} \in (1, 2)$. We use

$$\int (\mathrm{RG}_2(\boldsymbol{v}) - 2\tau\max(\boldsymbol{v}) + 2(\tau)^2 u)^2 du = \frac{(\mathrm{RG}_2(\boldsymbol{v}) - 2\tau\max(\boldsymbol{v}) + 2(\tau)^2 u)^3}{6(\tau)^2}.$$

We start with the case $\max(\boldsymbol{v}) \leq \tau$.

$$
\begin{aligned}
\mathrm{VAR}[\hat{\mathrm{RG}}_2^{(U)}|\boldsymbol{v}] &= (1 - \frac{\mathrm{RG}(\boldsymbol{v})}{\tau})\mathrm{RG}_4(\boldsymbol{v}) + \frac{(\mathrm{RG}_2(\boldsymbol{v}) - 2\tau\max(\boldsymbol{v}) + 2(\tau)^2 u)^3}{6(\tau)^2}\Big|_{\frac{\min(\boldsymbol{v})}{\tau}}^{\frac{\max(\boldsymbol{v})}{\tau}} \\
&= (1 - \frac{\mathrm{RG}(\boldsymbol{v})}{\tau})\mathrm{RG}_4(\boldsymbol{v}) + \frac{\mathrm{RG}_6(\boldsymbol{v})}{6(\tau)^2} - \frac{\mathrm{RG}_3(\boldsymbol{v})(\mathrm{RG}(\boldsymbol{v}) - 2\tau)^3}{6(\tau)^2} \\
&= \mathrm{RG}_3(\boldsymbol{v})(\frac{4}{3}\tau - \mathrm{RG}(\boldsymbol{v}))
\end{aligned}
$$

19

The case $\max(\boldsymbol{v}) \geq 2\tau$:

$$
\begin{aligned}
\text{VAR}[\hat{\text{RG}}_2^{(U)}|\boldsymbol{v}] &= (1 - \frac{\min(\boldsymbol{v})}{\tau})(\max(\boldsymbol{v})^2 - \text{RG}(\boldsymbol{v})_2)^2 + \frac{\min(\boldsymbol{v})}{\tau}(\max(\boldsymbol{v})^2 - 2\tau\max(\boldsymbol{v}) + \min(\boldsymbol{v})\tau - \text{RG}_2(\boldsymbol{v}))^2 \\
&= (\max(\boldsymbol{v}) - \text{RG}(\boldsymbol{v}))^2(\max(\boldsymbol{v}) + \text{RG}(\boldsymbol{v}))^2(1 - \frac{\min(\boldsymbol{v})}{\tau}) + \frac{\min(\boldsymbol{v})}{\tau}(\tau - \min(\boldsymbol{v}))^2(2\max(\boldsymbol{v}) - \min(\boldsymbol{v}))^2 \\
&= \frac{\tau - \min(\boldsymbol{v})}{\tau}\min(\boldsymbol{v})^2(2\max(\boldsymbol{v}) - \min(\boldsymbol{v}))^2 + \frac{\min(\boldsymbol{v})}{\tau}(\tau - \min(\boldsymbol{v}))^2(2\max(\boldsymbol{v}) - \min(\boldsymbol{v}))^2 \\
&= (2\max(\boldsymbol{v}) - \min(\boldsymbol{v}))^2\min(\boldsymbol{v})(\tau - \min(\boldsymbol{v}))
\end{aligned}
$$

We next handle the case $\tau \leq \max(\boldsymbol{v}) \leq 2\tau$, $\frac{\min(\boldsymbol{v})}{\tau} > \eta_0$.

$$
\begin{aligned}
\text{VAR}[\hat{\text{RG}}_2^{(U)}|\boldsymbol{v}] &= (1 - \frac{\min(\boldsymbol{v})}{\tau})(4\tau(\max(\boldsymbol{v}) - \tau) - \text{RG}_2(\boldsymbol{v}))^2 \\
&\quad + \frac{\min(\boldsymbol{v})}{\tau}\left(\frac{\tau\text{RG}_2(\boldsymbol{v})}{\min(\boldsymbol{v})} - 4\tau(\max(\boldsymbol{v}) - \tau)\frac{\tau - \min(\boldsymbol{v})}{\min(\boldsymbol{v})} - \text{RG}_2(\boldsymbol{v})\right)^2 \\
&= \frac{2(\tau - \min(\boldsymbol{v}))}{\tau}\left(\text{RG}_2(\boldsymbol{v}) - 4\tau(\max(\boldsymbol{v}) - \tau)\right)^2
\end{aligned}
$$

Lastly, for the case $\tau \leq \max(\boldsymbol{v}) \leq 2\tau$, $\frac{\min(\boldsymbol{v})}{\tau} \leq \eta_0$.

$$
\begin{aligned}
\text{VAR}[\hat{\text{RG}}_2^{(U)}|\boldsymbol{v}] &= (1 - \eta_0)\left(4\tau(\max(\boldsymbol{v}) - \tau) - \text{RG}_2(\boldsymbol{v})\right)^2 + \int_{\frac{\min(\boldsymbol{v})}{\tau}}^{\eta_0}(2(\tau(\max(\boldsymbol{v}) - u\tau) - \text{RG}_2)^2 du + \frac{\min(\boldsymbol{v})}{\tau}\text{RG}_4(\boldsymbol{v}) \\
&= \frac{\max(\boldsymbol{v}) - \tau}{\tau}\left(4\tau(\max(\boldsymbol{v}) - \tau) - \text{RG}_2(\boldsymbol{v}))^2 + \frac{\min(\boldsymbol{v})}{\tau}\text{RG}_4(\boldsymbol{v})\right. \\
&\quad + \frac{(\text{RG}_2(\boldsymbol{v}) + 4(\tau)^2 - 4\max(\boldsymbol{v})\tau)^3 - \text{RG}_3(\boldsymbol{v})(\text{RG}(\boldsymbol{v}) - 2\tau)^3}{6(\tau)^2}
\end{aligned}
$$

C Variance of $\hat{\text{RG}}^{(L)}$ and $\hat{\text{RG}}_2^{(L)}$

The estimators $\hat{\text{RG}}^{(L)}$ and $\hat{\text{RG}}_2^{(L)}$, provided in Tables 6 and 7 are obtained using (10). We calculate their variance.

Variance of $\hat{\text{RG}}^{(L)}$: When $\max(\boldsymbol{v}) \leq \tau$, we have $\hat{\text{RG}}^{(L)} = \tau\ln(\frac{\max(\boldsymbol{v})}{\tau\rho})$ when $1 \leq |S| \leq r - 1$ and $\hat{\text{RG}}^{(L)} = \tau\ln(\frac{\max(\boldsymbol{v})}{\min(\boldsymbol{v})})$ when $|S| = r$. The variance is

$$
\begin{aligned}
\text{VAR}[\hat{\text{RG}}^{(L)}|\boldsymbol{v}] &= (1 - \frac{\max(\boldsymbol{v})}{\tau})\text{RG}(\boldsymbol{v})^2 + \int_{\frac{\min(\boldsymbol{v})}{\tau}}^{\frac{\max(\boldsymbol{v})}{\tau}}(\text{RG}(\boldsymbol{v}) - \tau\ln(\frac{\max(\boldsymbol{v})}{\tau y}))^2 dy + \frac{\min(\boldsymbol{v})}{\tau}(\text{RG}(\boldsymbol{v}) - \tau\ln(\frac{\max(\boldsymbol{v})}{\min(\boldsymbol{v})})^2 \\
&= -2\tau\min(\boldsymbol{v})\ln(\frac{\max(\boldsymbol{v})}{\min(\boldsymbol{v})}) + 2\text{RG}(\boldsymbol{v})\tau - (\text{RG}(\boldsymbol{v}))^2
\end{aligned}
$$

When $\min(\boldsymbol{v}) \leq \tau \leq \max(\boldsymbol{v})$, $\hat{\text{RG}}^{(L)} = \max(\boldsymbol{v}) - \tau + \tau\ln(\frac{1}{\rho})$ when $1 \leq |S| \leq r - 1$ and $\hat{\text{RG}}^{(L)} = \max(\boldsymbol{v}) - \tau + \tau\ln(\frac{\tau}{\min(\boldsymbol{v})})$ when $|S| = r$. The variance is $\text{VAR}[\hat{\text{RG}}^{(L)}|\boldsymbol{v}] = (\tau)^2 - \min(\boldsymbol{v})^2 - 2\tau\min(\boldsymbol{v})\ln(\frac{\tau}{\min(\boldsymbol{v})})$.

Variance of $\hat{\text{RG}}_2^{(L)}$:

If $\min(\boldsymbol{v}) < \tau \leq \max(\boldsymbol{v})$, $\hat{\text{RG}}_2^{(L)} = \max(\boldsymbol{v})^2 - (\tau)^2 + 2\tau(u\tau - \max(\boldsymbol{v}) + \max(\boldsymbol{v})\ln\frac{1}{u}$ when $|S| \in [r - 1]$ and

$\hat{\text{RG}}_2^{(L)} = \max(\boldsymbol{v})^2 - (\tau)^2 + 2\tau(\min(\boldsymbol{v}) - \max(\boldsymbol{v}) + \max(\boldsymbol{v})\ln\frac{\tau}{\max(\boldsymbol{v})}$ when $|S| = r$. The variance is

$$\text{VAR}[\hat{\text{RG}}_2^{(L)}|\boldsymbol{v}] = 4\max(\boldsymbol{v})\min(\boldsymbol{v})\tau(\min(\boldsymbol{v}) - 2\max(\boldsymbol{v}))\ln\frac{\tau}{\min(\boldsymbol{v})}$$

$$+4\max(\boldsymbol{v})\min(\boldsymbol{v})(\tau)^2 + \frac{(\tau)^4}{3} - 6\max(\boldsymbol{v})\min(\boldsymbol{v})^2\tau - \min(\boldsymbol{v})^4$$

$$-4\max(\boldsymbol{v})^2\min(\boldsymbol{v})^2 + 4\max(\boldsymbol{v})\min(\boldsymbol{v})^3 + 4\max(\boldsymbol{v})^2(\tau)^2$$

$$-2\max(\boldsymbol{v})(\tau)^3 + \frac{8\min(\boldsymbol{v})^3\tau}{3}$$

If $\max(\boldsymbol{v}) < \tau$, $\hat{\text{RG}}_2^{(L)} = 2\tau(u\tau - \max(\boldsymbol{v}) + \max(\boldsymbol{v})\ln\frac{\max(\boldsymbol{v})}{u\tau})$ when $|S| \in [r-1]$ and $\hat{\text{RG}}_2^{(L)} = 2\tau(\min(\boldsymbol{v}) - \max(\boldsymbol{v}) + \max(\boldsymbol{v})\ln\frac{\max(\boldsymbol{v})}{\min(\boldsymbol{v})})$ when $|S| = r$. The variance is

$$\text{VAR}[\hat{\text{RG}}_2^{(L)}|\boldsymbol{v}] \quad = \quad -4\tau\max(\boldsymbol{v})\min(\boldsymbol{v})\ln(\frac{\max(\boldsymbol{v})}{\min(\boldsymbol{v})})(2\max(\boldsymbol{v}) - \min(\boldsymbol{v})) +$$

$$+\frac{2\tau}{3}(5\max(\boldsymbol{v})^3 - 9\max(\boldsymbol{v})\min(\boldsymbol{v})^2 + 4\min(\boldsymbol{v})^3) - \text{RG}_4(\boldsymbol{v})$$